Medicines from Microbes

Exploiting the power of computational genomics for natural products discovery and engineering

Marnix H. Medema

Cover design by Wart Burggraaf, ColouredPeople.com

Printed by Off Page, Amsterdam, NL. www.offpage.nl

ISBN: 978-90-367-6405-6 (printed version)

ISBN: 978-90-367-6406-3 (digital version)

The research described in this thesis was carried out at the Groningen Biomolecular Sciences and Biotechnology Institute (GBB), Department of Microbial Physiology and Groningen Bioinformatics Centre, University of Groningen, the Netherlands.

This research is supported by the Dutch Technology Foundation (STW), which is part of the Netherlands Organisation for Scientific Research (NWO) and partly funded by the Ministry of Economic Affairs (STW 10463).

This thesis was printed with additional financial support from DSM, the University of Groningen and the Groningen Graduate School of Science.

RIJKSUNIVERSITEIT GRONINGEN

Medicines from Microbes

Exploiting the power of computational genomics for natural products discovery and engineering

Proefschrift

ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
vrijdag 27 september 2013
om 11.00 uur

door

Marinus Hendrik Medema

geboren op 24 januari 1986
te Epe

Promotores:	Prof. dr. E. Takano Prof. dr. R. Breitling Prof. dr. L. Dijkhuizen
Beoordelingscommissie:	Prof. dr. J. Piel Prof. dr. P. Leadlay Prof. dr. M. Reinders

…from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved.

Charles Darwin, The Origin of Species

For from the greatness and beauty of created things comes a corresponding perception of their Creator.

Wisdom 13:5

The future indeed looks bright for the harvesting of useful genes, as whole genome sequencing becomes faster and cheaper and with the increasing ability to predict, just from a series of As, Cs, Gs and Ts, the small molecules likely to be made by the enzymes encoded by the gene clusters.

David Hopwood, J Ind Microbiol Biotechnol (2003) 30: 468–471

Table of Contents

Thesis abstract

Bacteria and fungi are capable of synthesizing a wide range of complex secondary metabolites. This thesis offers novel ways to develop drugs from these metabolites, by harnessing the power of bioinformatic computation and synthetic biology to drastically change the way in which new molecules are discovered and engineered. The thesis starts out with a general introduction (Chapter 1), followed by two distinct parts.

The first part describes a range of new software tools for the analysis of secondary metabolism. The computational tool antiSMASH (Chapter 2 and 3) automatically detects gene clusters that encode secondary metabolite biosynthetic pathways, and integrates this with substrate specificity predictions, gene annotations and comparative genomic analyses (which are expanded on in Chapter 4). Two additional software tools, Pep2Path (Chapter 5) and MultiMetEval (Chapter 6), allow the integration of genomic data on biosynthetic pathways with mass spectrometry and flux balance analysis, respectively. Finally, Chapter 7 outlines how these and related computational tools can be employed for the synthetic biology design of entirely novel biochemical pathways.

The second part describes the genomic analysis of secondary metabolite biosynthetic pathways and outlines strategies to utilize these pathways with methods offered by synthetic biology. Chapter 8 describes the genomic analysis of an extraordinary 'factory of antibiotics,' *Streptomyces clavuligerus*. Subsequently (in Chapter 9), a transcriptomic and metabolic modeling analysis is offered of an industrial strain of this species that has been used for the production of clavulanic acid. In Chapter 10, the horizon is expanded to the entire prokaryotic tree of life. The global analysis of secondary metabolism described here identifies an astonishing abundance of thousands of biosynthetic gene clusters in microbial genomes, and shows how nature continuously invents new molecular complexity by modular evolution of these gene clusters. In Chapter 11 and 12, a synthetic biology approach is outlined to exploit this large biochemical diversity by refactoring these gene clusters and inserting them in pre-optimized hosts using a plug-and-play strategy.

Finally, the concluding Chapter 13 outlines how bioinformatics will play a pivotal role in future research on secondary metabolism, by allowing effective integration of genomics, microbial ecology, high-throughput experimentation and synthetic biology.

Chapter 1

Mining microbial genomes for new secondary metabolites: a general introduction

Microbial cells are highly complex entities that interconvert hundreds of metabolites in order to survive, grow and replicate. Besides primary metabolites, involved in energy acquisition, cellular maintenance and growth, many microorganisms also produce *secondary metabolites*: these are substances that are non-essential for normal growth, but usually have important ecological functions. Secondary metabolites can, e.g., function as signaling molecules for cell–cell communication, as antibiotics in biochemical warfare, as siderophores for the harvesting of sparse nutrients, or as protective agents against radiation.

For humans, secondary metabolites are an important source of drugs, including antibiotics, immunosuppressants and antitumor agents, as well as food additives. In this context, they are also called *natural products*. Bacteria in general, and actinomycetes in particular, are the major source of antibiotics that are used in human and veterinary medicine (Hopwood 2007). Given the rapid emergence of antibiotic resistance in hospitals worldwide, microbial secondary metabolites have caught the renewed attention of scientists and pharmaceutical companies, as they may provide new weapons to combat multi-drug-resistant superbugs (Baltz 2008; Clardy, Fischbach, Walsh 2006; Fischbach and Walsh 2009).

Secondary metabolites: diverse molecules of distinct chemical classes

Secondary metabolites come in all shapes and sizes, and comprise a range of different classes of molecules (**Figure 1**). One of the broadest and best known classes of secondary metabolites are the peptides: small polymers of amino acids, often with one or more chemical modifications. Some peptides are translated by the ribosome directly from a small prepeptide-encoding mRNA. These ribosomally synthesized and post-translationally modified peptides (RiPPs) can be post-translationally modified in different ways, leading to various subclasses such as lantipeptides, thiopeptides, cyanobactins and microcins (Arnison et al. 2013). Famous examples of RiPPs are nisin, which is used as a food preservative, and the antibiotic thiostrepton.

Peptidic secondary metabolites are not only produced through translation by the ribosome. They can also be synthesized in a ribosome-independent fashion, through the action of large enzymes called nonribosomal peptide synthetases (Schwarzer, Finking, Marahiel 2003). These multi-domain enzymes can reach enormous sizes. They form large assembly lines, composed of distinct modules, each incorporating a specific amino acid into the growing peptide chain (Fischbach and Walsh 2006). Because these enzymes incorporate amino acids independent of the genetic code, nonribosomal peptides can contain a range of nonproteinogenic amino acids, in addition to the twenty ‘standard’ amino acids that are found in proteins. Several nonribosomal peptides, such as the antibiotics penicillin and vancomycin, the immunosuppressant cyclosporine and the cytostatic bleomycin, are widely used as medicines worldwide.

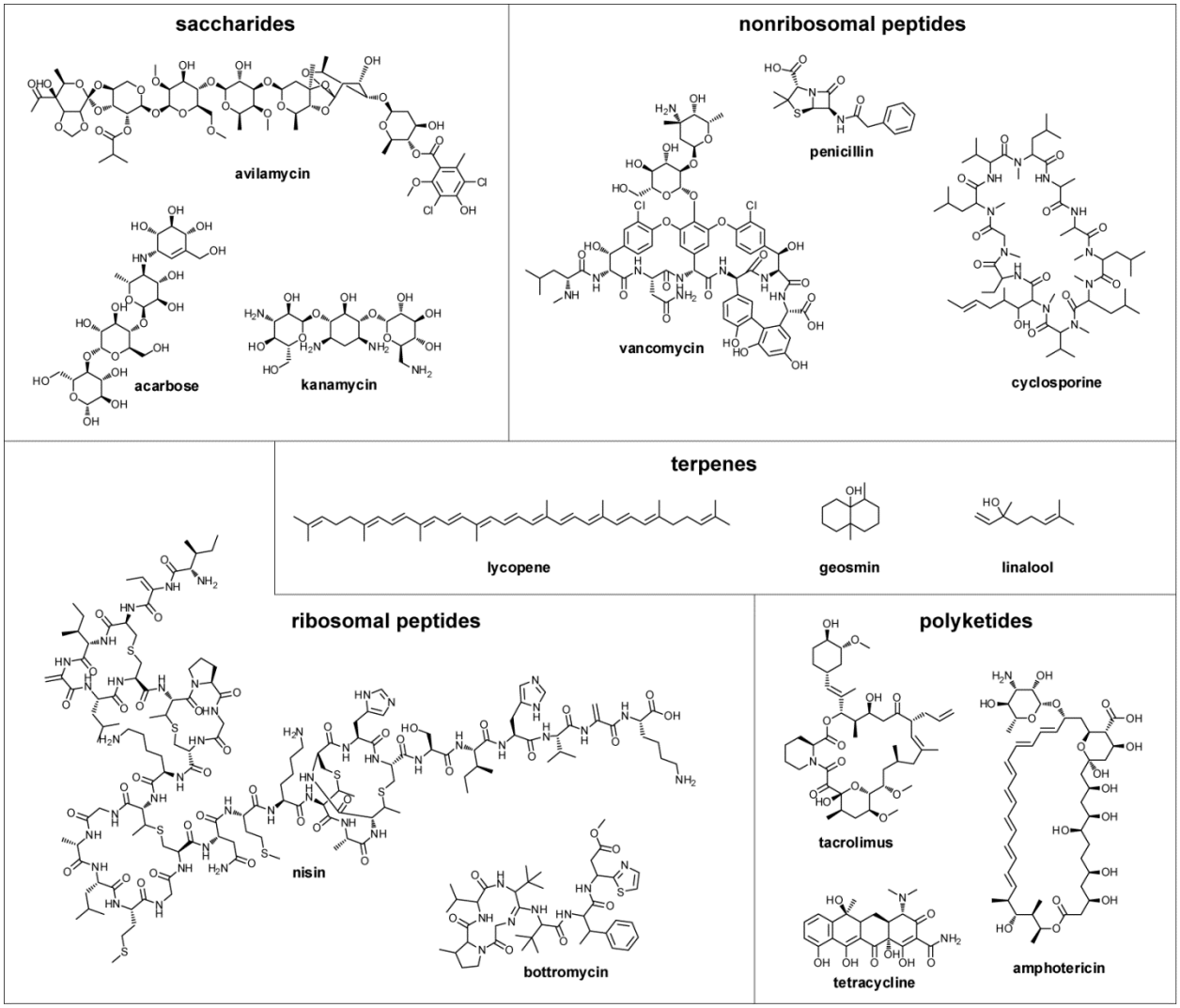

Figure 1. Examples of secondary metabolites that have been utilized by human society. Penicillin (Bentley 2005), tetracycline (Chopra, Hawkey, Hinton 1992), vancomycin (Barna and Williams 1984), kanamycin (Umezawa et al. 1957), avilamycin (Mertz et al. 1986) and bottromycin (Waisvisz et al. 1957) are antibiotics, amphotericin (Brajtburg et al. 1990) is an antifungal agent, cyclosporine (Graham 1994) and tacrolimus (Thomson, Bonham, Zeevi 1995) are immunosuppressants, acarbose (Wehmeier and Piepersberg 2004) is an anti-diabetic drug, lycopene (McDermott et al. 1973) is a red pigment, linalool (Nakano, Kim, Ohnishi 2011) and geosmin (Gerber and Lechevalier 1965) are aroma compounds and nisin (Delves-Broughton 1990) is a food preservative.

Another key class of secondary metabolites is formed by the polyketides, β-ketoacetyl polymers that comprise famous molecules such as the immunosuppressant tacrolimus, the antifungal drug amphotericin and the antibiotics tetracycline and erythromycin. Polyketides are synthesized by at least three different mechanisms. Similar to nonribosomal peptides, so-called 'type I' polyketides are also synthesized by multimodular enzymatic assembly lines (Fischbach and Walsh 2006). Besides the type I polyketides, there are also type II and type III polyketides, which are synthesized by separate single-domain enzymes (Shen 2003). Yet another large class of secondary metabolites are the terpenes, which are biosynthetically derived from isoprene units (Withers and Keasling 2007). They include the chemotherapeutic taxol, the flavoring agent menthol, and pigments like lycopene and β-carotene (which

make tomatoes red and carrots orange, respectively). The final major class of secondary metabolites is formed by saccharides (Musser 2003), which are small sugar polymers. These include the antidiabetic acarbose and the antibiotics avilamycin and kanamycin. Besides these five major classes, there exists a wide range of additional smaller classes of secondary metabolites, such as aminocoumarins, non-peptidic siderophores, indolocarbazoles, nucleosides, phenazines, carbapenems, butyrolactones, furans, and homoserine lactones (see (Walsh and Fischbach 2010) for a detailed review).

Figure 2. Schematic view of a (mock) biosynthetic gene cluster. The encoded molecular functions include biosynthesis of the core scaffold, tailoring/modification, precursor biosynthesis transport, self-resistance, and regulation.

Genomically, the biosynthetic pathways that produce secondary metabolites are usually encoded by groups of genes that are physically clustered on the chromosome. These *biosynthetic gene clusters* (BGCs) can range in size from a few kilobases (kb) to several hundred kb and may consist of several operons (transcriptional units). Often, they not only encode the enzymes that synthesize and/or modify the produced compound, but also enzymes that synthesize essential precursor molecules, membrane proteins that transport the end product out of the cell, proteins that confer resistance to the host if the molecule has antibiotic activity, and regulatory proteins that coordinate the gene expression of the transcriptional units in the BGC (**Figure 2**). The gene expression of BGCs is usually tightly regulated by complex regulatory mechanisms that respond to nutrient availability, the growth stage of the cells, the presence/absence of other organisms, or various causes of cellular stress (Bibb 2005; Martin and Liras 2010).

The impact of genomics

When the genome of the actinomycete bacterium *Streptomyces coelicolor* A3(2) was published in 2002, the authors identified more than twenty different BGCs coding for the production of secondary metabolites, even though only six of these metabolites were known at the time (Bentley et al. 2002). Subsequent genome sequencing of the bacteria *Streptomyces avermitilis* (Ikeda et al. 2003), *Pseudomonas fluorescens* (Paulsen et al. 2005), *Sorangium cellulosum* (Schneiker et al. 2007), and *Salinispora tropica* (Udwary et al. 2007) as well as the fungi *Aspergillus oryzae* (Machida et al. 2005), *Aspergillus fumigatus* (Nierman et al. 2005) and *Penicillium chrysogenum* (van den Berg et al. 2008) lead to the discovery of many dozens of BGCs for which no products were known. Intriguingly, there was only very little overlap in the BGCs present in the genomes of these species, suggesting that sequencing of

more microbial genomes could lead to the discovery of large numbers of novel secondary metabolites. Also, an explanation presented itself for the fact that many of these BGCs had gone unnoticed before genome sequencing: many of them appeared to be transcriptionally silent under typical laboratory conditions (or the product was synthesized in undetectably low amounts); they were therefore called 'cryptic' gene clusters (Scherlach and Hertweck 2009; Silakowski, Kunze, Müller 2001). Several techniques have now successfully been developed to awaken these cryptic gene clusters, such as the altering of growth conditions (Zazopoulos et al. 2003) and regulatory engineering of the genes controlling BGC expression (Gottelt et al. 2010).

The recent emergence of the new field of synthetic biology offers even more radical opportunities to mine these cryptic gene clusters for the discovery of novel pharmaceuticals: bottom-up genetic engineering with reusable parts could potentially facilitate the heterologous expression of BGCs from any source in a suitable host organism. By moving from the mere modification of existing DNA sequences to truly design-based engineering, all regulatory elements that hinder successful metabolite production can be replaced by synthetic ones that have been tested thoroughly and standardized for function under specific well-controlled conditions (Temme, Zhao, Voigt 2012).

Because of the rapid development of next-generation DNA sequencing technologies (Shendure and Ji 2008) and the recent sequencing of thousands of microbial genomes (Pagani et al. 2012), colossal numbers of cryptic BGCs are now becoming available for study, offering great potential to discover many novel antibiotics, anti-tumor agents and other useful compounds. Traditional approaches to natural product discovery no longer suffice. To fully exploit this potential, it is vital that computational methodologies are developed to detect, compare and prioritize BGCs, and that synthetic biology strategies are designed to express and screen the most promising BGCs in high-throughput. This thesis aims to offer both.

Outline of the thesis

The thesis consists of two parts. Part I presents a range of new software tools for the genomic analysis of secondary metabolism, whereas Part II applies these and other tools to identify and compare secondary metabolite biosynthesis pathways across the microbial tree of life.

In the first two chapters of Part I (**Chapter 2 and 3**), antiSMASH is introduced: a computational pipeline and web server for the rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. antiSMASH not only allows users to detect up to 24 different classes of BGCs, it also uses several substrate specificity prediction algorithms to predict the approximate chemical structure of the compound encoded by a BGC, based on the encoded enzyme amino acid sequences. Moreover, it compares every BGC with a database of BGCs detected by the algorithm in all available genome sequences. The underlying comparative gene cluster analysis algorithm was also made into a more versatile stand-alone tool: MultiGeneBlast (**Chapter 4**). MultiGeneBlast allows the user-friendly comparison of gene clusters or gene cluster fragments of any size with either the entire GenBank database or with user-customized databases designed from local

files, GenBank entries or combinations thereof. **Chapter 5** discusses Pep2Path, an algorithm that harnesses the power of antiSMASH to directly link its results to mass spectrometry data, in order to discover novel bioactive peptides in high-throughput and link them to their corresponding BGCs. Coupling the power of genomic analysis to constraint-based metabolic modeling, **Chapter 6** introduces MultiMetEval, an intuitive tool for comparative analysis of genome-scale metabolic models. It shows how this new type of analysis can be applied to predict theoretical flux maxima of secondary metabolite biosynthesis pathways for multiple bacterial hosts, which can guide the engineering and heterologous expression of BGCs with synthetic biology approaches. Finally, **Chapter 7** highlights how multiple software tools may be combined to mine biosynthetic pathways and engineer them with the tools offered by synthetic biology.

Part II starts off with the analysis of the genome and transcriptome of *Streptomyces clavuligerus* ATCC 27064, an industrially important bacterium that produces multiple antibiotics as well as clavulanic acid, a beta-lactam inhibitor. In **Chapter 8**, I show that the genome of *S. clavuligerus* contains an unprecedented number of BGCs, about half of which are located on an enormous linear plasmid. **Chapter 9** then provides a transcriptomic analysis of an industrial clavulanic acid-overproducing strain obtained by multiple rounds of random mutagenesis, which offers clues on how genomic mutations can cause rerouting of metabolic fluxes towards the production of a specific compound. **Chapter 10** expands the view from a single organism of high interest to the entire microbial tree of life, and shows the result of an attempt to globally and systematically identify all BGCs in bacterial and archaeal genome sequences. I show how these data can be exploited to pinpoint unexplored regions in the 'biosynthetic universe,' and how this leads to the identification of hundreds of BGCs that encode compounds that appear to fall entirely outside known chemical classes of secondary metabolites. The chapter also sheds light on the evolution of BGCs, showing how they often evolve by the insertion/deletion of 1–10 kb multigene modules, how BGCs are horizontally transferred even over great taxonomic distances, and how concerted evolution often homogenizes DNA sequences of genes encoding repetitive multi-domain proteins, obscuring the mutual evolutionary relationships between BGCs. In order to fully exploit all the treasures identified in all the thousands of available genome sequences, **Chapter 11** proposes a synthetic biology plug-and-play strategy to re-engineer and express BGCs in pre-optimized cellular systems. **Chapter 12** then delineates this strategy in more detail, explaining how BGCs can be computationally prioritized and then experimentally refactored in practice. Finally, **Chapter 13** offers conclusions and future perspectives on how computational genomics and synthetic biology will reshape the field of microbial secondary metabolism.

Part I

Software tools for the genomic analysis and design of secondary metabolic pathways

Chapter 2

antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences

Published as:

M.H. Medema, K. Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research* 39: W339-W346.

Abstract

Bacterial and fungal secondary metabolism is a rich source of novel bioactive compounds with potential pharmaceutical applications as antibiotics, anti-tumor drugs or cholesterol-lowering drugs. To find new drug candidates, microbiologists are increasingly relying on sequencing genomes of a wide variety of microbes. However, rapidly and reliably pinpointing all the potential gene clusters for secondary metabolites in dozens of newly sequenced genomes has been extremely challenging, due to their biochemical heterogeneity, the presence of unknown enzymes and the dispersed nature of the necessary specialized bioinformatics tools and resources. Here, we present antiSMASH (antibiotics & Secondary Metabolite Analysis Shell), the first comprehensive pipeline capable of identifying biosynthetic loci covering the whole range of known secondary metabolite compound classes (polyketides, nonribosomal peptides, terpenes, aminoglycosides, aminocoumarins, indolocarbazoles, lantibiotics, bacteriocins, nucleosides, beta-lactams, butyrolactones, siderophores, melanins and others). It aligns the identified regions at the gene cluster level to their nearest relatives from a database containing all other known gene clusters, and integrates or cross-links all previously available secondary-metabolite specific gene analysis methods in one interactive view. antiSMASH is available at http://antismash.secondarymetabolites.org.

The website is free and open to all and there is no login requirement.

Introduction

Microbial secondary metabolites offer great potential for the development of new medicines. They belong to a wide variety of chemical classes, and many of them have cholesterol-lowering, anti-tumor or antibiotic activities. The rapid decrease in the cost of genome sequencing now allows the discovery of hundreds or even thousands of gene clusters encoding the biosynthetic machinery for these compounds (Walsh and Fischbach 2010). However, laboratory research cannot keep pace with the speed of genomic discovery, as the experimental characterization of each gene cluster is still very laborious. Therefore, effective *in silico* identification of the most promising targets within genomes is essential for the successful mining of the genomic riches available. Manual annotation is very labor-intensive and time-consuming, leading to incomplete annotations. Automatic annotation of secondary metabolite clusters with subsequent manual curation may enhance accuracy as well as completeness of the annotation. A few *in silico* methods have been published thus far to automate the analysis of secondary metabolism in bacterial genomes. The first of these was ClustScan (Starcevic et al. 2008), which allows the uploading of genomic data to a server for the semi-automatic detection and annotation of polyketide synthase (PKS) and nonribosomal peptide synthetase (NRPS) gene clusters. Additionally, Anand et al. (2010) recently published the SBSPKS toolbox for structure-based PKS analysis. Li et al. (2009) constructed the NP.searcher web server, which is specialized in predicting the possible chemical structures resulting from a subset of gene cluster types. Unfortunately, all these tools are largely limited to the analysis of the core genes for type I polyketide (PK) and nonribosomal peptide (NRP) biosynthesis. Thus far, accessory genes as well as core genes for many other secondary metabolite scaffolds have largely been

neglected in computational approaches, even though some very good but also very specific tools are available for bacteriocin (de Jong et al. 2010) and type III PKS (Mallika et al. 2010) detection. For fungal genomes, the SMURF tool (Khaldi et al. 2010) has recently become available, which is capable of generating a somewhat more comprehensive list of secondary metabolite biosynthesis gene clusters, but this tool offers little further detailed analysis. Thus far, CLUSEAN (Weber et al. 2009) offered the most comprehensive analysis by including a full genome annotation, but is difficult to operate for the nonspecialist and requires intensive manual analysis of the output.

Here, we present a software pipeline for secondary metabolite gene cluster identification, annotation and analysis which is comprehensive, rapid and user-friendly (**Figure 1**). It can be run either from a web server (http://antismash.secondarymetabolites.org/) or as a stand-alone version on a standard desktop computer. It can rapidly detect all known classes of secondary metabolite biosynthesis gene clusters, provide detailed NRPS / PKS functional annotation, and predict the chemical structure of NRPS/PKS products with higher accuracy than existing methods.

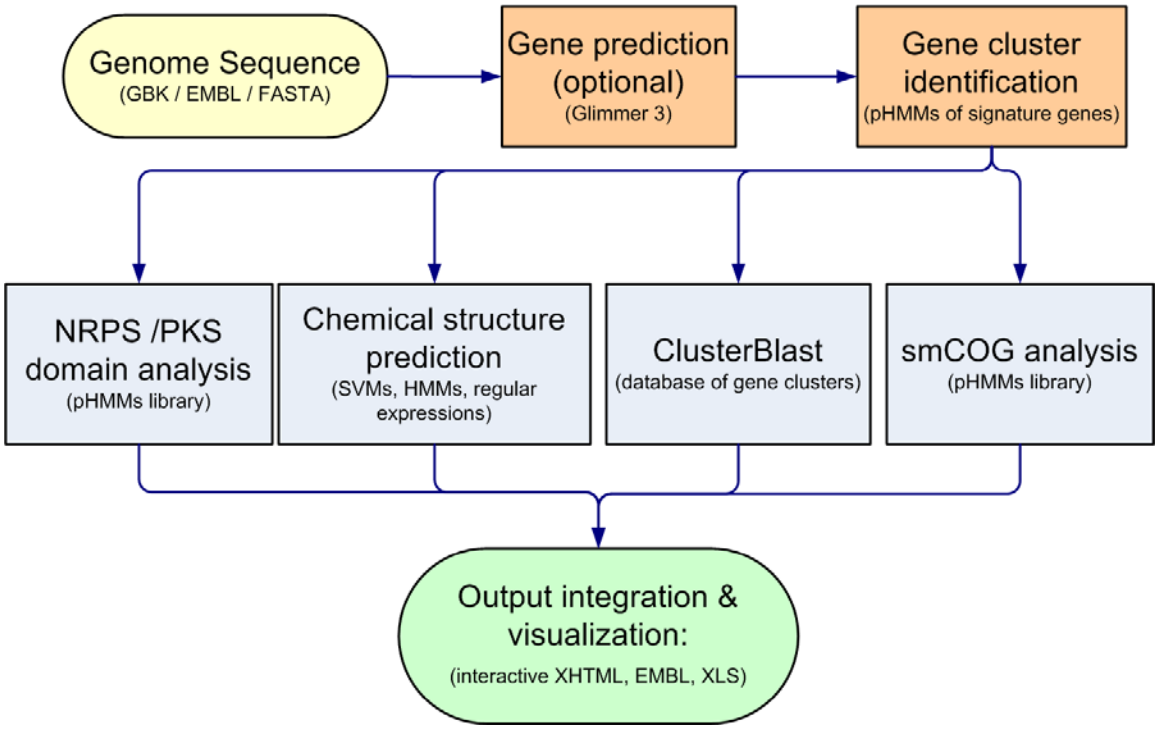

Figure 1: Outline of the pipeline for genomic analysis of secondary metabolites. Genes are extracted or predicted from the input nucleotide sequence, and gene clusters are identified with signature gene pHMMs. Subsequently, several downstream analyses can be performed: NRPS/PKS domain analysis and annotation, prediction of the core chemical structure of PKSs and NRPSs, ClusterBlast gene cluster comparative analysis, and smCOG secondary metabolism protein family analysis. The output is visualized in an interactive XHTML web page, and all details are stored in an EMBL file for additional analysis and editing in a genome browser. A Microsoft Excel file with an overview of all detected gene clusters and their details is also generated.

Additionally, by constructing a database of all currently known secondary metabolite biosynthesis gene clusters throughout the tree of life, we were able to equip the tool with a comparative gene cluster analysis module. In this module, evolutionary similarities between a queried gene cluster and other gene clusters are detected and visualized in order to be able to rapidly infer functions of genes and operons based on homology. Finally, from the genes within this database of gene clusters, we constructed secondary metabolism Clusters of Orthologous Groups (smCOGs). These are used in yet another module to predict and categorize the functions of accessory genes, and to calculate phylogenetic trees for each gene with a seed alignment of its smCOG protein family. Our benchmark results show that our method reliably detects gene clusters of a wide variety of biosynthetic types, and that it is able to significantly enhance manual genome annotations of secondary metabolite biosynthesis.

Methods and Implementation

File and options input

The input front end of the antiSMASH web server allows uploading of sequence files of a variety of types (FASTA, GBK, or EMBL files). Alternatively, a GenBank/RefSeq accession number can be provided, which is used by the web server to automatically obtain the associated file from GenBank. If the user chooses to use a FASTA input file, gene prediction is performed by Glimmer3 (Delcher et al. 2007) — using its long-orfs tool to construct a gene model based on the input sequence itself — or by GlimmerHMM (Majoros, Pertea, Salzberg 2004) when eukaryotic input data is submitted. Before starting the antiSMASH analysis run, the user can select the gene cluster types he or she wants to search for. Additionally, he or she can select which of the downstream analysis modules to include. For those users who, e.g., work with proprietary data, a stand-alone version with a Java graphical user interface is available with the same input options as the web version. Finally, expert users may choose to directly run the Python-based pipeline program from the command line in order to batch-analyze a larger number of inputs.

Detection of secondary metabolite biosynthesis gene clusters

Using the HMMer3 tool (http://hmmer.janelia.org/), the amino acid sequence translations of all protein-encoding genes are searched with profile Hidden Markov Models (pHMMs) that are based on multiple sequence alignments of experimentally characterized signature proteins or protein domains (proteins, protein subtypes or protein domains which are each exclusively present in a certain type of biosynthetic gene clusters). Using both existing pHMMs (de Jong et al. 2010; Finn et al. 2010; Letunic, Doerks, Bork 2009; Yadav, Gokhale, Mohanty 2009) and new pHMMs from seed alignments, we constructed a library of models specific for type I, II and III PK, NRP, terpene, lantibiotic, bacteriocin, aminoglycoside / aminocyclitol, beta-lactam, aminocoumarin, indole, butyrolactone, ectoine, siderophore, phosphoglycolipid, melanin, and aminoglycoside biosynthesis signature genes. Additionally, we

constructed a number of pHMMs specific for false positives, such as the different types of fatty acid synthases which show homology to PKSs. The final detection stage operates a filtering logic of negative and positive pHMMs and their cut-offs. The logic is based on knowledge of the minimal core components of each gene cluster type taken from the scientific literature. The cut-offs were determined by manual studies of the pHMM results when run against the NCBI non-redundant protein sequence (nr) database (ftp://ftp.ncbi.nlm.nih.gov/blast/db). All technical details on the pHMM library and the detection rules are available in **Supplementary Tables I** and **II**, respectively.

Gene clusters are defined by locating clusters of signature gene pHMM hits spaced within <10 kilobases (kb) mutual distance. To include flanking accessory genes, gene clusters are extended by 5, 10, or 20 kb on each side of the last signature gene pHMM hit, depending on the gene cluster type detected. As a consequence of this greedy methodology, gene clusters that are spaced very closely together may be merged into 'superclusters'. These gene clusters are indicated in the output as 'hybrid clusters'; they may either represent a single gene cluster which produces a hybrid compound that combines two or more chemical scaffold types, or they may represent two separate gene clusters which just happen to be spaced very closely together.

NRPS / PKS domain architecture analysis

NRPS / PKS domain architectures are analyzed (**Figure 2**) using another pHMM library, which contains existing models (Ansari et al. 2008; Finn et al. 2010; Letunic, Doerks, Bork 2009; Rausch et al. 2007; Weber et al. 2009; Yadav, Gokhale, Mohanty 2009) as well as newly constructed models specific for NRPS/PKS protein domains and functional / phylogenetic subgroups of these domains (**Supplementary Table III**).

Conserved motifs within key PKS and NRPS domains are also detected using the pHMMs described earlier in the CLUSEAN package (Weber et al. 2009), and are written to the detailed downloadable EMBL output. PKS/NRPS gene names are annotated according to the domains and domain subtypes that the genes contain (e.g. "hybrid NRPS-PKS", "enediyne PKS", "glycopeptide NRPS", "trans-AT PKS", etc.).

Substrate specificity, stereochemistry and final structure predictions

Substrate specificity prediction of PKS and NRPS modules, based on the active sites of their respective acyltransferase (AT) and adenylation (A) domains, is performed by various available methods. PKS AT domain specificities are predicted using a twenty-four amino acid signature sequence of the active site (Yadav, Gokhale, Mohanty 2003), as well as with pHMMs based on the method of Minowa et al. (2007), which is also used to predict co-enzyme A ligase domain specificities. NRPS A domain specificities are predicted using both the signature sequence method and the support-vector machines-based method of NRPSPredictor2 (Rausch et al. 2005; Röttig et al. 2011), and using the method of Minowa et al. (2007). Finally, all predictions are integrated into a consensus prediction by a majority vote. Ketoreductase

domain-based stereochemistry predictions for PKSs (Starcevic et al. 2008) are performed as well. An estimate of the biosynthetic order of PKS / NRPS modules is predicted based on PKS docking domain sequence residue matching (for type I modular PKSs, (Anand et al. 2010)) or assumed colinearity, and a final predicted core chemical structure is generated as a SMILES string (Weininger 1988), i.e. a unique text description of the chemical structure, and visualized in a picture file (**Figure 2**). To increase the reliability of the core structure prediction, monomers for which there was no consensus in the predictions are represented as generic amino acids or ketides with unspecified R-groups.

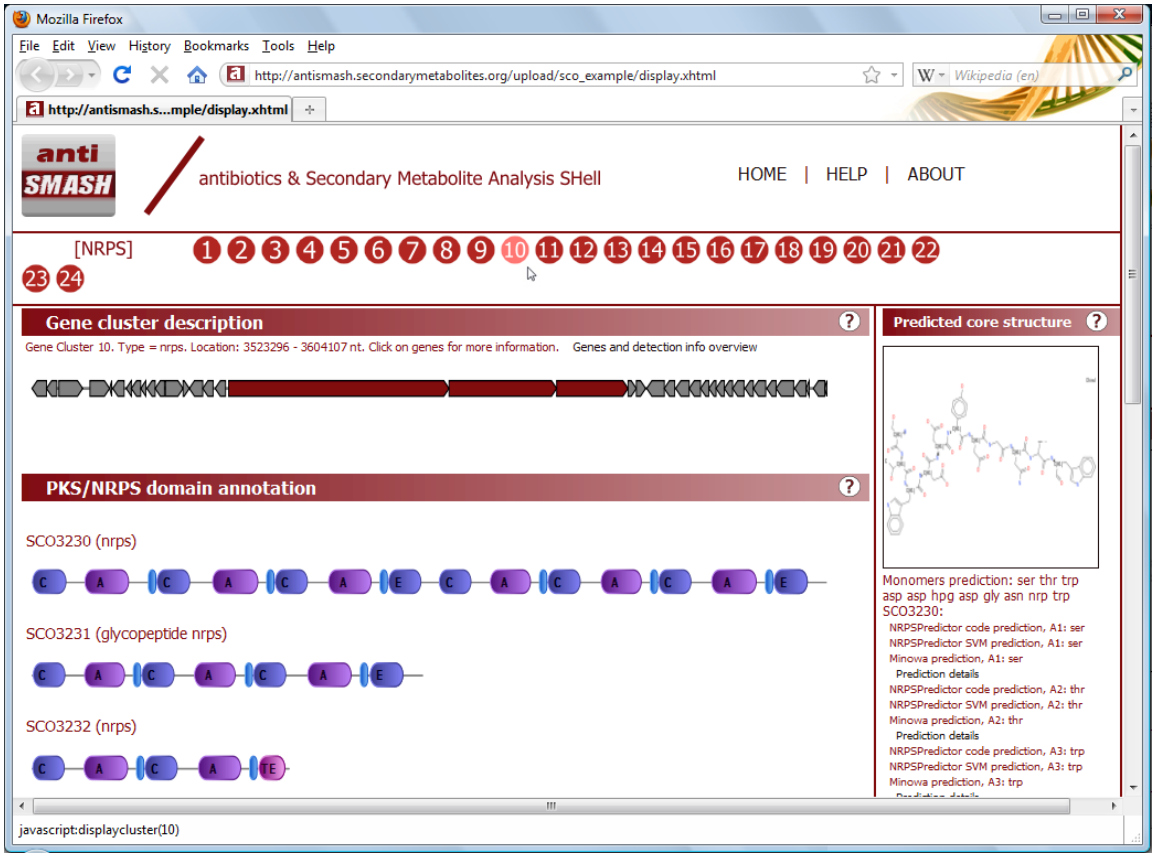

Figure 2: Interactive XHTML visualization of results. The numbers below the banner represent the gene clusters that were detected, the type of which is shown to the left of them at mouse-over. Once a gene cluster has been selected, the 'Gene cluster description' tab will display an SVG image with all genes within the approximate gene cluster, with the detected signature genes displayed in red. Locus tags appear on mouse-over, and on clicking a gene a small panel pops up with annotation information and cross-links to other web services. If PKS/NRPS proteins are encoded in the gene cluster, their domain annotations are given in the 'PKS/NRPS domain annotation' tab. More detailed domain annotation information and cross-links are provided on mouse-over. In the 'Predicted core structure' tab, a prediction of the core chemical structure is given for PKS or NRPS gene clusters based on the predictions displayed below it. All tabs contain a wide range of links to pop-ups which further detail the prediction information.

Secondary metabolite Clusters of Orthologous Groups

In order to rapidly annotate the accessory genes surrounding the detected core signature genes in the various types of secondary metabolite biosynthesis gene clusters, we constructed a database of all gene clusters contained in the latest NCBI nt database (15/02/2011). To do so, pHMMs described above were used to detect all secondary metabolite biosynthesis gene cluster signature genes in the nr database. The accession numbers of all hits meeting the described cut-offs were extracted and used to download the corresponding GenPept files. If the taxonomy identifier included "bacteria" or "fungi", the nucleotide source accession number was extracted. The corresponding nucleotide GenBank files were then downloaded as well, and cross-checked for presence of the queried protein accession number. For each nucleotide GenBank file, gene clusters were detected as described above. Amino acid sequences of all genes contained within the gene clusters were written to a FASTA file with headers containing key information, and a summary of all detected gene clusters (nt accession, nt description, cluster number, cluster type, protein accession numbers) was written to a text file. To construct the smCOGs, clustering of all gene cluster proteins was performed using OrthoMCL (Li, Stoeckert, Roos 2003), and consensus annotations were manually assigned based on the frequencies of the five most prevalent annotations of each smCOG in GenBank. For each smCOG, a seed alignment was created from 100 randomly picked sequences using MUSCLE 3.5 (Edgar 2004), and a pHMM of each smCOG was generated based on the conserved core of each alignment (**Supplementary Figure 1**). Within the antiSMASH software pipeline, the smCOG pHMMs are used for functional annotation of all accessory genes within the gene clusters. After assignment of an smCOG to a gene — based on the highest-scoring pHMM on its sequence above a certain e-value treshold — the predicted protein sequence is aligned to the smCOG seed alignment, and a rough neighbor-joining phylogenetic tree is calculated using FastTree 2 (Price, Dehal, Arkin 2010) and visualized with TreeGraph 2 (Stover and Müller 2010) (**Supplementary Figure** 1).

ClusterBlast comparative gene cluster analysis

Secondary metabolite biosynthesis gene clusters are highly modular, and their genes are transferred frequently from one gene cluster to another during evolution (Donadio et al. 2005; Fischbach, Walsh, Clardy 2008). Therefore, when trying to obtain a functional understanding of a gene cluster, it is highly beneficial to be able to compare it with (parts of) other gene clusters which show similarity to it and which may have been characterized experimentally. In order to facilitate this, we applied our annotated database of gene clusters to link up protein sequences with their parent gene clusters and create a comparison tool — based on the most recent BLAST+ implementation (Camacho et al. 2009) — which ranks gene clusters by similarity to a queried gene cluster. Clusters are sorted first based on an empirical similarity score $S = h + H + s + S + B$, in which h is the number of query genes with a significant hit, H is the number of core query genes with a significant hit, s is the number of gene pairs with conserved synteny, S is the number of gene pairs with conserved synteny involving a core gene, and B is a core gene bonus (3 points given when at least one core gene has a hit in the subject cluster). If the similarity scores are equal, the hits are subsequently ranked based on the cumulative BlastP bit scores between the gene clusters. This feature enables a rapid assessment of the comparative genomics for each annotated cluster (**Figure 3**).

Figure 3: Example of ClusterBlast alignment of gene clusters homologous to the query gene cluster. In this case, the ten best hits to the calcium-dependent antibiotic NRPS gene cluster from *Streptomyces coelicolor* A3(2) are displayed. Homologous genes (BLAST e-value < 1E-05; 30% minimal sequence identity; shortest BLAST alignment covers over >25% of the sequence) are given the same colors. The 'select gene cluster alignment' drop-down menu provides links to one-by-one gene cluster alignments to each gene cluster hit. In the one-by-one gene cluster alignments, PubMed and/or PubChem links are provided for gene clusters associated with a known compound.

Genome-wide BLAST and Pfam analysis and prediction of potential unknown secondary metabolite biosynthesis gene cluster types

To facilitate further thorough manual genome analysis, antiSMASH has also been linked up to the whole-genome BLAST and Pfam analysis modules from the previously published CLUSEAN framework (Weber et al. 2009). the CLUSEAN results are integrated into an EMBL output file. Furthermore, as unknown biosynthetic gene cluster types are likely to exist which may be missed by the antiSMASH gene cluster detection module, the Pfam results are also used to predict genomic regions with a high probability of constituting secondary metabolite biosynthesis gene clusters in a more generalized fashion than the signature genes pHMMs method. For this, the genome sequence is converted to a

string of predicted Pfam domains which is fed to a hidden Markov model (Cimermancic et al., in preparation) with transitions between a gene cluster state and a rest-of-the-genome state. This model was trained on Pfam domain frequencies from a set of 473 cloned gene clusters (gene cluster state) and from the set of ~1100 genomes currently in the JGI IMG database (rest-of-the-genome state). The result of this analysis is visualized in a PNG graph.

Output and visualization

All pipeline analysis results are visualized in a user-friendly interactive XHTML page (**Figure 2**), which can be used to browse through the different gene clusters. For PKS and NRPS gene clusters, the predicted core chemical structures are shown as images. Gene cluster maps are drawn with scalable vector graphics (SVGs), to which interactive on-click and mouse-over functions are added through JavaScript to provide annotation information, pipeline result scores, and BLAST hyperlinks. Detected signature genes on which the gene cluster identification is based are shown in a distinct color. ClusterBlast results are displayed in a similar way, as aligned gene cluster maps in which genes with mutual BLAST hits are given identical colors. Additionally, available at the bottom right of the page, fully annotated EMBL output files provide the user with the additional possibility to browse their genome in a genome browser such as Artemis (Rutherford et al. 2000).

Results

Compared to previous software, the pipeline described here is uniquely comprehensive: it integrates all previously published analysis types into one tool and adds valuable novel functionalities (**Table I**).

Software	Open-source & stand-alone available	Covers full tree of life	NRPS/PKS detection	NRPS/PKS detailed functional domain annotation	NRP / PK core structure prediction	Detection of other biosynthetic classes	Gene cluster border prediction	Comparative gene cluster analysis	Prediction of all secondary metabolite-like genomic regions
ClustScan		+	+	+	+	±			
CLUSEAN	+		+	+					
NP.searcher	+	+	+		+				
SBSPKS		+	+	+					
SMURF			+			±	+		
antiSMASH	+	+	+	+	+	+	+	+	+

Table I: Comparison of different software tools for secondary metabolite biosynthesis analysis. Comparison of functionalities of currently existing programs or software packages for secondary metabolite biosynthesis analysis.

In order to measure the accuracy of the gene cluster predictions, we performed two independent benchmark evaluations of the method. Firstly, we collected the sequences of cloned gene clusters of known compounds of biosynthetic types by searching both the GenBank / RefSeq databases and the scientific literature with a range of different keywords. From the resulting set of 484 cloned gene cluster GenBank files, 473 (97.7%) were correctly identified by antiSMASH, and 468 (96.7%) were given exactly the same annotation by antiSMASH as by the articles describing their experimental characterization (**Figure 4**, **Supplementary Table IV**).

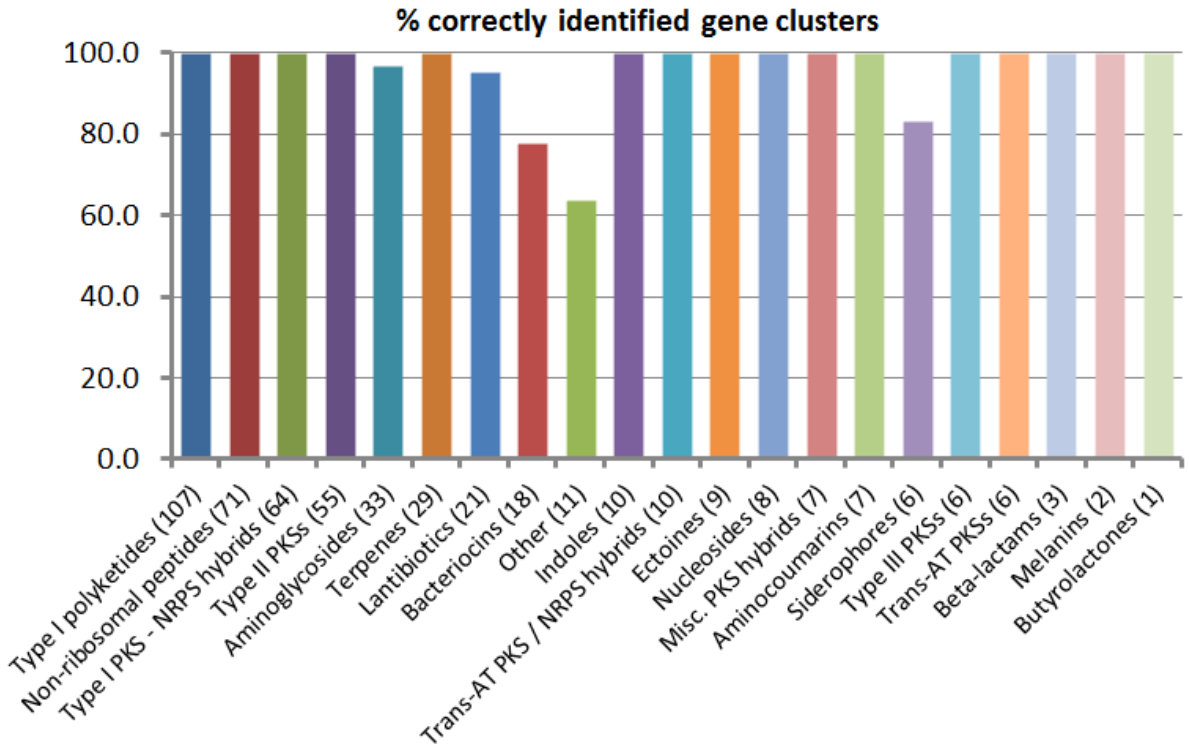

Figure 4: Benchmark results on a set of 473 cloned secondary metabolite biosynthesis gene clusters found in the GenBank nucleotide database. The numbers behind the names of the biosynthetic types indicate how many gene clusters of that type were in the benchmark set.

In order to test for false positives as well, we also benchmarked the method on five well-annotated genomes from different taxonomic groups. Besides genomes of three different actinomycetes (the organisms on which the tool is likely to be used most often) these included a proteobacterium (*Pseudomonas fluorescens* Pf-5) and a fungus (*Aspergillus fumigatus* Af293). In the five genomes, 97.3% of all 111 annotated gene clusters were detected by antiSMASH (**Figure 5**, **Supplementary Table V**). Under closer scrutiny, two of the three gene clusters that were missed by antiSMASH appeared to lack a complete set of genes associated with biosynthesis of a known chemical scaffold. More interestingly, 35 additional gene clusters were detected (31.5%) which had been missed during initial genome annotation and which after close inspection all appeared to have a high probability of being actual biosynthetic gene clusters. The cluster types that appeared to be frequently missed during the annotation of these genomes appeared to be butyrolactones (eight gene clusters missed), terpenes (seven gene clusters missed), NRPSs/PKSs (six gene clusters missed) and lantibiotics (five gene clusters missed), which

suggests that the computational approach used can yield improvements even in finding gene clusters of common biosynthetic types.

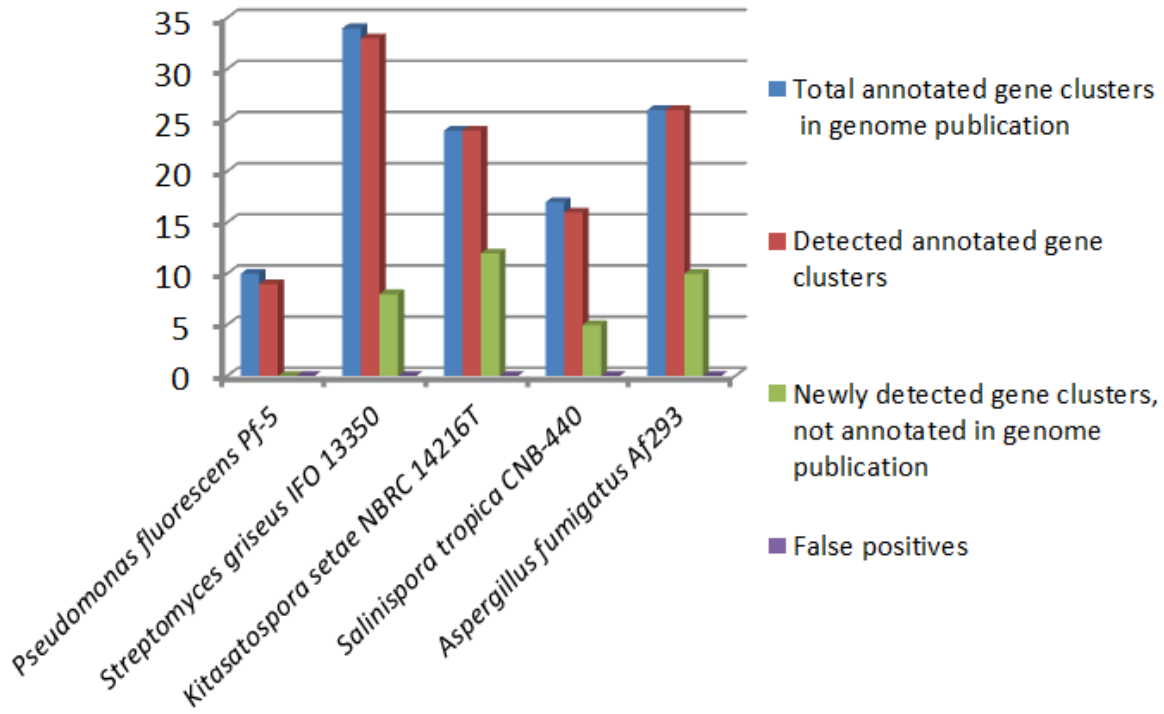

Figure 5: Benchmark results on five genome sequences. All except three annotated gene clusters from the five genome publications were detected; two of these annotated gene clusters (SGR5285-SGR5295 in *Streptomyces griseus* and Strop_3244-Strop_3253 in *Salinispora tropica*) appeared to lack core genes for biosynthesis of a known secondary metabolite chemical scaffold. The one certain gene cluster which was not detected was a small gene cluster for the biosynthesis of hydrogen cyanide from *Pseudomonas fluorescens* Pf-5.

We also compared the performance of antiSMASH with other existing tools. No similarly comprehensive tools are available, but NP.searcher and SMURF each offer automated gene cluster detection for a small subset of the cluster types detected by antiSMASH (NP.searcher detects bacterial NRPS/PKS gene clusters, and SMURF detects fungal NRPS, PKS, and dimethylallyl tryptophan synthase gene clusters). Our analysis of the results of these tools on four bacterial and two fungal genomes (**Supplementary Table VI**), respectively, showed that antiSMASH and SMURF performed equally well (both detect 74 gene clusters, with 93.4% overlap). Compared to NP.searcher, antiSMASH detected significantly more (47 vs. 31, i.e. 51.6% more) NRPS/PKS gene clusters, while all NP.searcher-detected gene clusters were also picked up by antiSMASH. The gene clusters that were detected by antiSMASH but not by NP.searcher were all small NRPS-like or PKS-like gene clusters. None of the three tools gave predictions that were clear false positives, except one SMURF detection of a probable fatty acid synthase (GenBank ID CAP98191.1) that was labeled as a PKS.

Discussion and Conclusions

antiSMASH not only provides a unique integration of previously widely dispersed tools, but it also achieves very high accuracy in its individual cluster annotations, which are enhanced by unique novel analyses such as BLAST-based gene cluster alignments and secondary metabolite COG phylogenetic trees for accessory genes. As the field of synthetic biology is opening up new ways to study these gene clusters in a high-throughput fashion (Medema et al. 2011a), antiSMASH will enable experimental researchers to quickly pinpoint those gene clusters most interesting for further study, and swiftly collect secondary metabolite BioBricks for the (re-)design of gene clusters. Moreover, the new comparative analyses that antiSMASH offers provide unprecedented possibilities to interpret the functions of both complete gene clusters and their particular genes in their evolutionary context. The approaches developed are likely to soon allow global analysis of all small molecule biosynthesis gene clusters throughout the tree of life, so that we can acquire a more and more comprehensive understanding of how nature itself designs novel bioactive compounds.

Funding

This work was supported by the Dutch Technology Foundation STW, which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs [STW 10463]; and by the GenBioCom program of the German Ministry of Education and Research (BMBF) [grant 0315585A]. RB is supported by an NWO-Vidi fellowship, and ET by a Rosalind Franklin Fellowship, University of Groningen. MHM was supported by a travel grant from the Boehringer Ingelheim Fonds.

Acknowledgements

We thank Mike Li for kindly providing a script for the conversion of strings of amino acid and/or polyketide residues into SMILES strings. We thank Marc Röttig and Oliver Kohlbacher for providing NRPSpredictor2.

Supplementary Material

Supplementary figure and tables can be downloaded from http://rdmy.info/ch2

Chapter 3

antiSMASH2 – a versatile platform for genome mining of secondary metabolite producers

Published as:

K. Blin*, M.H. Medema*, D. Kazempour, M.A. Fischbach, E. Takano, R. Breitling, T. Weber (2013) antiSMASH2 – a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Research* 41: W204-W212.

*Equal contribution

Abstract

Microbial secondary metabolites are a potent source of antibiotics and other pharmaceuticals. Genome mining of their biosynthetic gene clusters has become a key method to accelerate their identification and characterization. In 2011, we developed antiSMASH, a web-based analysis platform that automates this process. Here, we present the highly improved antiSMASH 2.0 release, available at http://antismash.secondarymetabolites.org/. For the new version, antiSMASH was entirely redesigned using a plug-and-play concept that allows easy integration of novel predictor or output modules. antiSMASH 2.0 now supports input of multiple related sequences simultaneously (multi-FASTA/GenBank/EMBL), which allows the analysis of draft genomes comprised of multiple contigs. Moreover, direct analysis of protein sequences is now possible. antiSMASH 2.0 has also been equipped with the capacity to detect additional classes of secondary metabolites, including oligosaccharide antibiotics, phenazines, thiopeptides, homo-serine lactones, phosphonates and furans. The analysis module for predicting the core structure of the cluster end product is now also covering lantipeptides, in addition to polyketides and non-ribosomal peptides. The antiSMASH ClusterBlast functionality has been extended to identify sub-clusters involved in the biosynthesis of specific chemical building blocks. All in all, the new features currently make antiSMASH 2.0 the most comprehensive resource for identifying and analyzing novel secondary metabolite biosynthetic pathways in microorganisms.

Introduction

Many microorganisms produce secondary metabolites with interesting bioactivities, many of which are applied as antibiotics, anti-cancer agents and many other drugs (Newman and Cragg 2012).

For decades, the only way to identify and characterize such bioactive secondary metabolites involved a labor- and time-consuming procedure: one had to isolate new bacterial or fungal strains, cultivate them under different conditions, identify, isolate, purify and test any bioactive molecules that were produced, and perform a complete chemical structure elucidation. The rapidly decreasing cost of whole-genome sequencing technologies enables new approaches that can greatly accelerate this process using bioinformatics analysis of the genome sequences of potential producer strains (Crawford and Clardy 2012; Scheffler et al. 2013; Zotchev, Sekurova, Katz 2012), prior to or in parallel with the biological/chemical isolation process. The fact that the biosynthetic pathways for many secondary metabolites are encoded by highly modular compact gene clusters facilitates this kind of analysis (Medema et al. 2011b; Medema et al. 2012).

In recent years, many individual algorithms have been developed which cover specific steps in the bioinformatics analysis of secondary metabolite biosynthesis based on microbial genome sequences (for review see (Fedorova, Moktali, Medema 2012; Weber 2013)). For example, ClustScan (Starcevic et al. 2008), CLUSEAN (Weber et al. 2009), SBSPKS (Anand et al. 2010), and SMURF (Khaldi et al. 2010) are tools for the identification and/or analysis of the enzymatic domains in multi-modular polyketide synthases and/or non-ribosomal peptide synthetases, which are the key enzymes for the synthesis of

Table I. Overview of the capabilities of various software tools for the analysis of biosynthetic gene clusters. antiSMASH 2.0 combines by far the most functionalities into a single framework, and adds four key new features compared to antiSMASH 1.0. The phylogenomic analysis embedded in NaPDoS can be accessed through direct links from the relevant C and KS domains shown in the antiSMASH output page.

Features	antiSMASH 2.0	antiSMASH 1.0	CLUSEAN	SMURF	ClustScan	NaPDoS	NP.searcher	NRPSpredictor2	NRPSSP	SBSPKS
Open-source & stand-alone available	X	X	X				X	X		X
Covers bacteria, archaea and fungi	X	X			X		X	X	X	X
NRPS/PKS detection	X	X	X	X	X	X	X	X	X	X
NRPS/PKS detailed functional domain annotation	X	X	X		X			X		X
NRP / PK core structure prediction	X	X			X		X		X	
Lantipeptide core structure prediction	X									
Detection of other biosynthetic classes	X	X		X						
Gene cluster border prediction	X	X		X						
Comparative gene cluster analysis	X	X								
Sub-cluster analysis	X									
Prediction of putative novel gene cluster types	X	X								
Protein sequence input	X					X		X	X	X
Nucleotide sequence input	X	X	X	X	X	X	X			
Multi-contig input	X					X				
PKS structural modeling										X
NRPS/PKS domain phylogenomic analysis	(X)[1]					X				

[1] Support for NRPS/PKS phylogenomic analysis via NaPDoS cross-reference

prominent classes of clinically important secondary metabolites. These include, e.g., non-ribosomal peptide antibiotics like penicillin and polyketide macrolides like the immunosuppressant tacrolimus. NRPSpredictor (Rausch et al. 2005; Röttig et al. 2011), NRPSsp (Prieto et al. 2012), and the PKS/NRPS predictive Blast Server (Bachmann and Ravel 2009) are sophisticated tools for the prediction of substrate specificities of key biosynthetic steps, allowing an approximate prediction of the chemical structure of bioactive end compounds based on the genome sequence (**Table I**).

In 2011, we released the first version of the "antibiotics and Secondary Metabolite Analysis SHell" (antiSMASH), a web server and stand-alone software, which combines automated identification of secondary metabolite gene clusters in genome sequences with a large collection of compound-specific analysis algorithms (Medema et al. 2011b). Within the last two years, antiSMASH has become the standard tool to analyze genomes of bacteria and fungi for their potential to produce secondary metabolites. Since the start of the service, the stand-alone software has been downloaded more than 3,200 times, and more than 28,000 antiSMASH jobs have been submitted to the antiSMASH webserver; the monthly data volume currently processed is over 12 gigabases. antiSMASH also supports the manual PKS/NRPS cluster curation effort of the ClusterMine360 database (Conway and Boddy 2013) by providing a standardized annotation basis.

Here, we present version 2.0 of antiSMASH. The software has been entirely restructured internally, and now utilizes a plug-and-play concept for easier maintainability and extensibility. A number of novel cluster detection and analysis features have been added to cover the broadest possible range of secondary metabolite classes. Finally, the web-based user interface was completely redesigned for better usability and a wider range of possible input files, allowing, e.g., the analysis of unassembled draft genomes and metagenomic sequences.

Methods and implementation of new features

The basic steps of an antiSMASH analysis have been described by Medema et al. (2011b): first, potential biosynthetic gene clusters are identified by comparing each gene product encoded on the uploaded DNA sequence against a manually curated collection of profile Hidden Markov Models (pHMMs). These pHMMs describe key biosynthetic enzymes of the 24 secondary metabolite classes detectable by antiSMASH, using the HMMer3 software (Eddy 2011). Key enzymes encoded in each gene cluster are assigned to secondary metabolite-specific Clusters of Orthologous Groups (smCOGs). Depending on the class of the detected secondary metabolite gene cluster, further detailed analyses are performed: the domains of multimodular polyketide synthases (PKSs) and non-ribosomal peptide synthetases (NRPSs) are identified by a pHMM-based approach. Specificities of enzymes are determined by analyzing active site residues using integrated third-party algorithms and tools, such as the methods of Minowa et al. (2007) and NRPSpredictor2 (Röttig et al. 2011) for the prediction of NRPS adenylation domain specificities. Based on these data, a core chemical structure of the putative biosynthesis product is generated and displayed. In addition, an integrated version of MultiGeneBlast (Medema, Takano, Breitling 2013), ClusterBlast, is used to identify similar gene clusters in a comprehensive gene cluster database. antiSMASH 2.0 can be either installed locally on Windows, Mac OS X, or Linux computers, or be accessed via the internet at http://antismash.secondarymetabolites.org (recommended). The use of the antiSMASH web server is free of charge, and does not require registration or login data. Voluntarily, the users can provide

an email address, which is used to send information and the link to the results, once the computing of the antiSMASH 2.0 results is finished. The data are stored on the server for 30 days and are deleted afterwards.

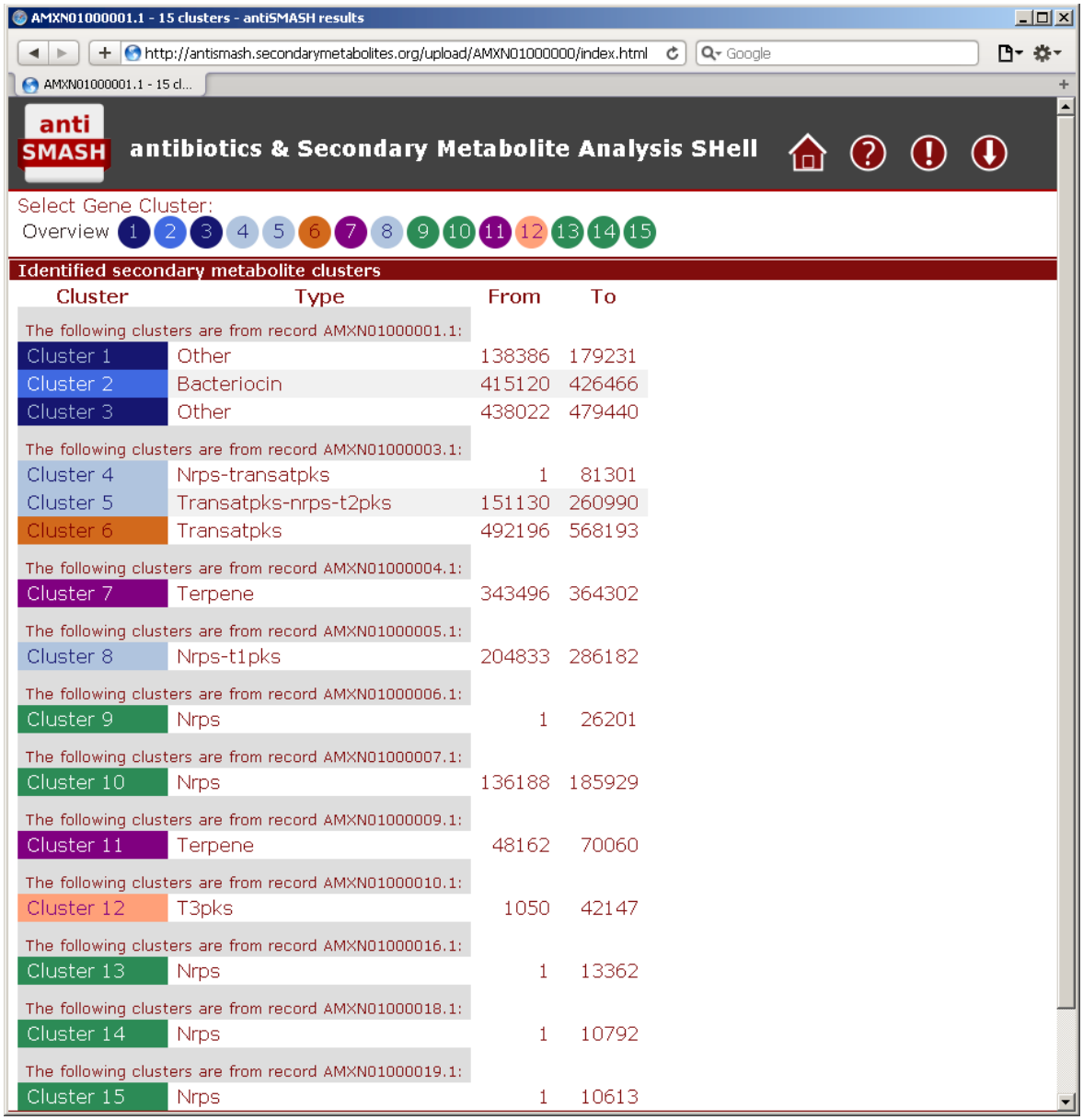

Figure 1. Overview page of the antiSMASH results. antiSMASH 2.0 gives an overview of all the output results in a single page, showing all the detected biosynthetic gene clusters with their type classifications and nucleotide positions. For inputs consisting of multiple entries / contigs, the clusters are separated by input entry / contig. Gene cluster types are signified by specific colors.

While the general strategy of antiSMASH has not changed in version 2.0, many improvements have been implemented in the new version, which are outlined below.

New file and input options

antiSMASH 2.0 now makes it easier to work with draft genomes consisting of a large number of individual sequence records: support has been added for multi-GenBank, multi-EMBL, as well as multi-FASTA files. If the NCBI download option yields a whole genome shotgun (WGS) master or supercontig record, antiSMASH 2.0 will download all constituent single WGS records from NCBI as well and combine all of them into a single output (**Figure 1**). For prokaryotic FASTA inputs, antiSMASH 2.0 now also offers the option to perform the initial search for gene cluster signature genes on all open reading frames of >60 nt throughout all six translation frames of a nucleotide sequence, before running the standard gene prediction with Glimmer. This avoids that mistakes in the gene prediction stage lead to false negatives in the gene cluster prediction stage. After the gene prediction stage, all open reading frames that match to pHMMs in the antiSMASH pHMM library are retained in the gene cluster output, even if they were not predicted as genes by Glimmer.

In addition to nucleotide sequences, antiSMASH 2.0 can now also be used to analyze PKS, NRPS and lantipeptide precursor amino acid sequences directly: their protein sequences can either be analyzed by specifying their NCBI GenPept accession numbers or by pasting the FASTA sequences directly into an input field.

Detection of secondary metabolite gene clusters in sequence data

In addition to the secondary metabolite cluster types supported in the original release of antiSMASH (type I, II and III polyketides, non-ribosomal peptides, terpenes, lantipeptides, bacteriocins, aminoglycosides / aminocyclitols, beta-lactams, aminocoumarins, indoles, butyrolactones, ectoines, siderophores, phosphoglycolipids, melanins, and a generic class of clusters containing unusual secondary metabolite biosynthesis genes), version 2.0 adds support for oligosaccharide antibiotics, phenazines, thiopeptides, homoserine lactones, phosphonates and furans. The cluster detection uses the same profile Hidden Markov Model (pHMM) rule-based approach as the initial release (Medema et al. 2011b): in short, the pHMMs are used to detect signature proteins or protein domains that are characteristic for the respective secondary metabolite biosynthetic pathway. Some pHMMs were obtained from PFAM or TIGRFAM. If no suitable pHMMs were available from these databases, custom pHMMs were constructed based on manually curated seed alignments (**Supplementary Table I**). These are composed of protein sequences of experimentally characterized biosynthetic enzymes described in literature, as well as their close homologs found in gene clusters from the same type. The models were curated by manually inspecting the output of searches against the non-redundant (nr) database of protein sequences. The seed alignments are available online at http://antismash.secondarymetabolites.org/download.html#extras. After scanning the genome with the pHMM library, antiSMASH evaluates all hits using a set of rules (**Supplementary Table II**) that describe the different cluster types. Unlike the hard-coded rules in the initial release of antiSMASH, the detection rules and profile lists are now located in editable TXT files, making it easy for users to add and modify cluster rules in the stand-alone version, e.g. to accommodate newly discovered or proprietary compound classes without code changes. The results of gene cluster predictions by antiSMASH are continuously checked on new data arising from research performed throughout the

natural products community, and pHMMs and their cut-offs are regularly updated when either false positives or false negatives become apparent.

The profile-based detection of secondary metabolite clusters has now been augmented by a tighter integration of the generalized PFAM (Punta et al. 2012) domain-based ClusterFinder algorithm (Cimermancic et al., in preparation) already included in version 1.0 of antiSMASH. This algorithm performs probabilistic inference of gene clusters by identifying genomic regions with unusually high frequencies of secondary metabolism-associated PFAM domains, and was designed to detect "classical" as well as less typical and even novel classes of secondary metabolite gene clusters. Whereas antiSMASH 1.0 only generated the output of this algorithm in a static image, version 2.0 displays these additional putative gene clusters along with the other gene clusters in the HTML output. A key advantage of this is that these putative gene clusters will now also be included in the subsequent (Sub)ClusterBlast analyses.

Metabolite-specific detection modules

antiSMASH version 2.0 adds lantipeptide-specific chemical core structure analysis to the existing set of NRPS/PKS core prediction tools. If one or more open reading frames encoding putative lantipeptide prepropeptides are found, antiSMASH predicts the core peptide molecular mass and sequence after leader peptide cleavage. The leader peptide cleavage motifs are identified via pHMMs specific for cleavage sites of class I to IV lantipeptides, respectively. The best-matching profile determines the classification of the prepropeptide, and the cleavage site is calculated from the pHMM–sequence alignment.

To obtain the core peptide mass, all serine and threonine residues in the core peptide are assumed to be dehydrated to didehydro-alanine (Dha) and didehydro-butyrine (Dhb), the most frequent post-translational modification in lantipeptides. Reported masses are the monoisotopic masses of the most prevalent isotopomers. The number of lanthionine/methyl-lantionine bridges is calculated from the number of cysteine, Dha and Dhb residues available for bridge formation (Blin et al., manuscript in preparation).

SubclusterBlast

Extending the ClusterBlast analysis that identifies homologous gene clusters across many published genome sequences, we have added a new option to identify operons or other sets of genes related to the biosynthesis of precursors or specific chemical moieties in a gene cluster's end product. This new analysis module, SubclusterBlast, performs blastp searches of the amino acid translations of all cluster genes against a database containing 126 sub-clusters from gene clusters encoding known compounds (**Figure 2**).

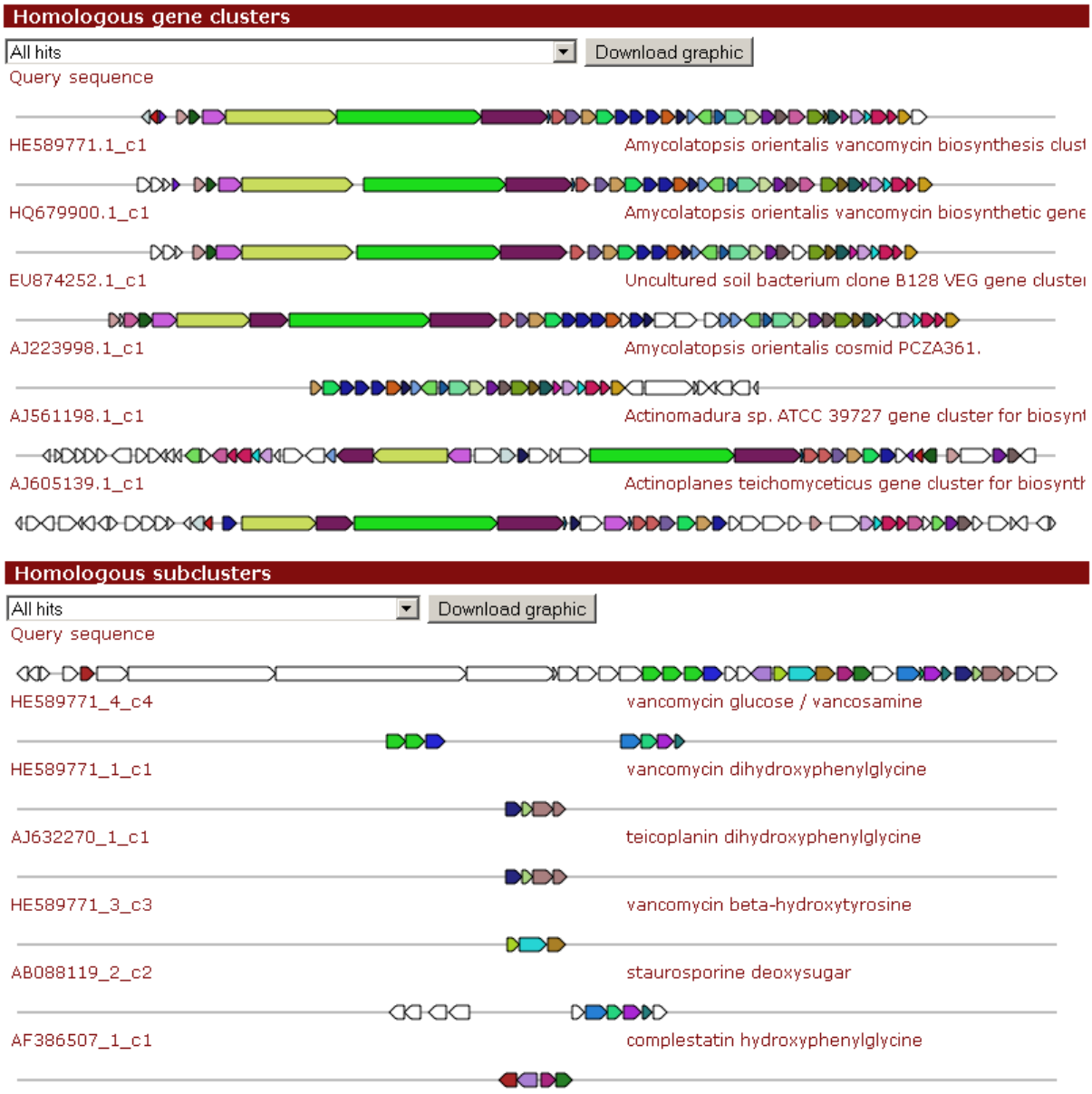

Figure 2. ClusterBlast and SubClusterBlast outputs for the balhimycin biosynthesis gene cluster. The top six hits of each analysis module are shown. The ClusterBlast module shows the homology between the balhimycin gene cluster (Pelzer et al. 1999) and the vancomycin, VEG, A40926, and teicoplanin biosynthesis gene clusters. Homologous genes are shown in identical colors, while white-colored genes have no blast hits between the gene clusters. The novel SubclusterBlast module can identify homologous sub-clusters encoding the biosynthesis of specific chemical moieties. In this case, SubclusterBlast is able to identify the dihydroxyphenylglycine (dHpg), hydroxyphenylglycine (Hpg) and hydroxytyrosine (Bht) precursor biosynthesis sub-clusters, as well as the vancosamine-like sugar biosynthesis sub-cluster.

These sub-clusters code for the biosynthesis of precursors such as 6-methylsalicylic acid, 3-amino-5-hydroxybenzoic acid, ethylmalonyl-CoA, deoxysugars and hydroxyphenylglycine, which are highly specific for certain classes of bioactive compounds. Hence, their presence in a genome allows more confident conclusions about the biosynthetic capacities of an organism. The hits are sorted in the same way as the ClusterBlast hits (Medema et al. 2011b), but are gathered with stricter thresholds: a minimal percentage identity of 45% and a minimal sequence coverage of 40% are required. The highest-scoring sub-cluster hits are then displayed on the results page using an annotated vector graphic similar to the general ClusterBlast output.

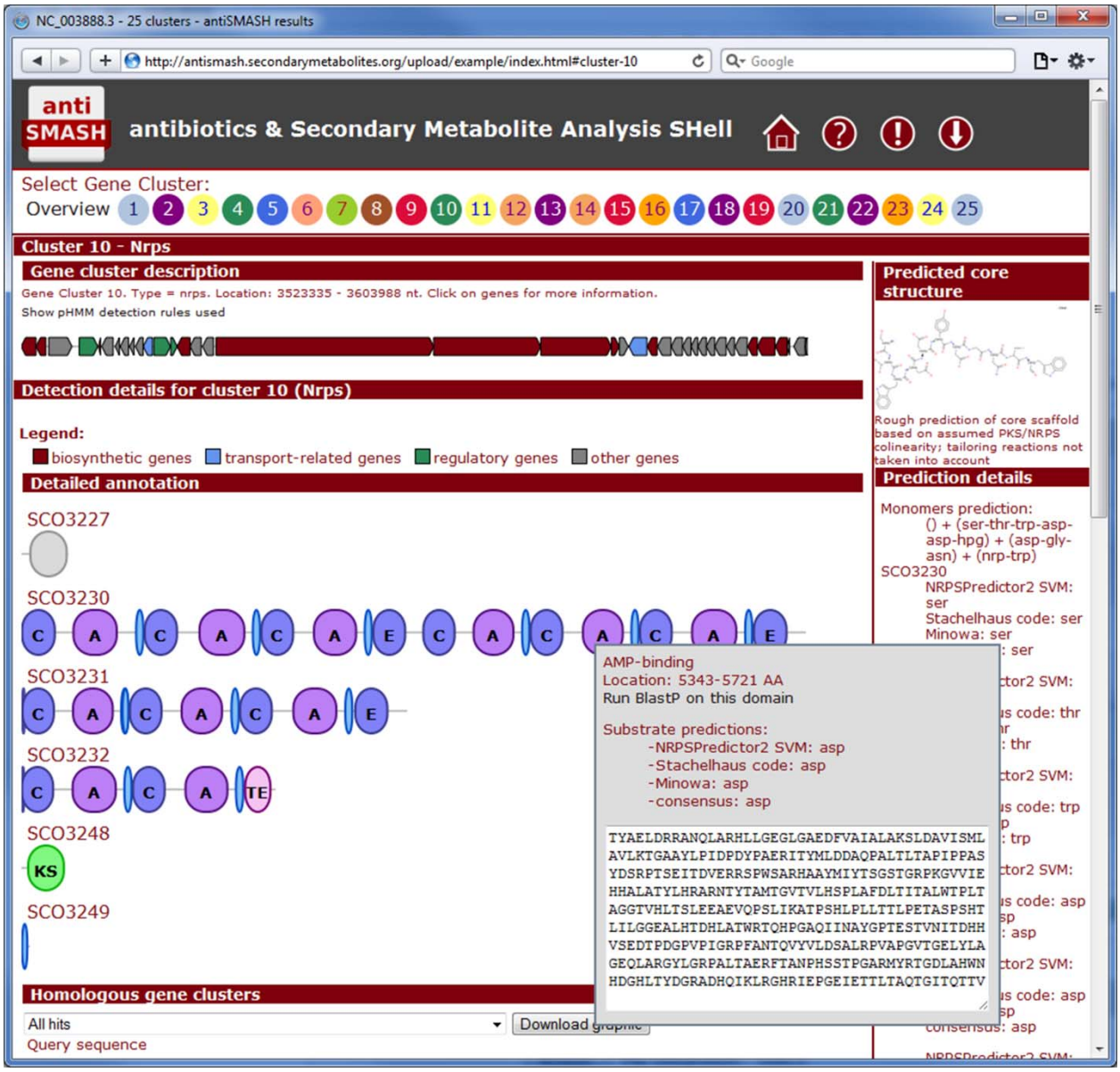

Figure 3. Top part of a gene cluster overview in the redesigned antiSMASH 2.0 output. The gene cluster shown is the calcium-dependent antibiotic biosynthesis gene cluster from *Streptomyces coelicolor* A3(2). The gene cluster type-specific coloring of the numbered gene cluster buttons makes it easier to navigate through large result files. smCOG-based coloring of biosynthetic, transport-related and regulatory genes within the gene cluster make it easier to interpret the architecture of the gene cluster.

Output and visualization

When antiSMASH has finished the computation of an analysis, it now provides an overview table that displays all identified secondary metabolite biosynthesis gene clusters with links to the respective prediction details, as a convenient starting point for further analysis (**Figure 1**). For nucleotide inputs consisting of multiple GBK/EMBL/FASTA entries, the results are separated per entry. Due to the large size of the antiSMASH results webpage in version 1.0, loading took quite a long time and sometimes even caused timeout error messages in the user's web browser. Therefore, the visualization component of antiSMASH 2.0 was completely redesigned, resulting in a reduction of transfer data volume and greatly accelerated display, even for results containing many cluster hits.

The overall layout of the interactive results page has been retained (**Figure 3**): in the top section, the identified clusters are displayed as circles that serve as direct links to the clusters. In antiSMASH 2.0, the circles are color-coded depending on the class of the identified cluster to ease navigation by the user. The individual cluster result pages are now reachable via the result URL, making it possible to both bookmark and direct other people to specific cluster pages. Individual cluster result pages contain an interactive graphical representation of the genes identified in the cluster. Again, color-coding was added to represent the functional classes of the gene cluster genes according to an smCOG-based classification: biosynthesis, transport, regulation, or other. For modular enzymes (NRPS, PKS) or lantipeptides, detailed annotation sections provide information on the domain organization and the putative cleavage sites and molecular weights, respectively. At the bottom of the page, graphical representations of the ClusterBlast results and – if available – the SubclusterBlast results are displayed. For several classes of secondary metabolites, where the analysis of the gene clusters allows the prediction of core structures of the biosynthetic products, a predicted structure and detailed information on the prediction source are displayed in a box on the right side of the results page **(Figure 3**). For lantipeptides and NRPS products, there is a direct link to the NORINE (Caboche et al. 2008) peptide database. The information displayed on the interactive webpage is also annotated in EMBL- or GenBank-formatted sequence files, which can be downloaded and used with standard sequence analysis software. In addition, an archive containing all data including the webpage can be saved for later use.

Plug-and-play architecture

In antiSMASH 2.0, the software architecture has been completely redesigned to make it easily extendable: the core program reads in "analysis plug-ins" that are either general or specific to a certain gene cluster type. "Output plug-ins" facilitate the output of the results to HTML, GBK, EMBL, TXT and XLS files. To make it easy for users to customize antiSMASH for their own analyses, we provide a plug-in template from the download section of http://antismash.secondarymetabolites.org, which can be used to design custom plug-ins, e.g. for reading user-specific input formats or analyzing novel cluster types.

Results and discussion

With options to upload DNA sequences of both finished genomes and draft sequences, to make antiSMASH download published sequences from NCBI, and to analyze amino acid sequences directly, antiSMASH 2.0 now covers all common types of input data. For draft genome data published in the NCBI genome database, antiSMASH can automatically download the records specified in the WGS summary record. As a test for the downloader, the recently published *Oxytricha trifallax* WGS record (Genbank accession no. AMCR00000000.1) consisting of 22,363 contigs was run via the internet interface, and the server handled the large amount of contigs and sequence data (67 megabases) without issues. For prokaryotic genome sequences, draft genome support increases the number of genomes that can be processed directly via NCBI accession numbers from 2570 to 8898, a ~2.5-fold increase of available sequences. One important caveat should be noted: when analyzing draft genomes, the number of detected gene clusters reported by antiSMASH can be artificially high,

because gene clusters can be fragmented across multiple contigs, and antiSMASH detects all fragments as separate gene clusters. On the other hand, some contigs with gene cluster fragments might be left undetected, if the subset of genes present on a contig does not suffice to match the criteria for gene cluster detection by antiSMASH.

antiSMASH 2.0 now supports 24 secondary metabolite cluster types via profile-based detection of their core biosynthetic genes (up from 19). In test runs on twenty-eight known gene clusters encoding compounds of the newly added classes, all of them were detected successfully (**Supplementary Table 3**). To assess the general accuracy of the antiSMASH predictions, we selected the same test set of genomes as for the original version (Medema et al. 2011b): The genomes of the proteobacterium *Pseudomonas fluorescens* Pf-5 (Paulsen et al. 2005), the actinomycetes *Streptomyces griseus* IFO 13350 (Ohnishi et al. 2008), *Kitasatospora setae* NBRC 14216T (Ichikawa et al. 2010), and *Salinispora tropica* CNB-440 (Udwary et al. 2007), and the fungus *Aspergillus fumigatus* Af293 (Nierman et al. 2005) were analyzed with antiSMASH 2.0 and compared to the manually identified clusters referred to in the original publications. 97.3 % of clusters (108 of 111) that were assigned manually were also identified by antiSMASH 2.0. This is the same performance as with antiSMASH 1.0, which was expected, as the established cluster finding algorithm has not changed in version 2.0. In addition to the 35 clusters which were predicted by antiSMASH 1.0 but were missed in the original publications, 4 additional clusters were identified by the new detection modules of antiSMASH 2.0, increasing the percentage of newly found gene clusters from 31.5% to 35.1% (**Supplementary Table 4**).

If further extension of the prediction ability is desired, new profiles can be added easily and without changes to the core code of the software using the new plug-and-play architecture of antiSMASH 2.0. antiSMASH 2.0 can also cast a wider net than the original version, by using improved ways to exploit the outputs of the ClusterFinder inclusive search algorithm for putative clusters (Cimermancic et al., in preparation). While the inclusive algorithm is likely to identify too many clusters, the combination with homology search methods allows focusing on the ones with homology to previously identified secondary metabolite clusters.

A major goal of antiSMASH 2.0 was to increase usability. Because antiSMASH 1.0 loaded all the results simultaneously when loading/opening the HTML output file, it was slow for the typical large results files: e.g., loading the 35 cluster results for *Streptomyces tsukubaensis* NRRL18488 (Genbank accession no. AJSZ01000001) from a local hard drive took around 40 seconds on a fast PC. In contrast, antiSMASH 2.0 output for the same data now loads in less than two seconds, even though more clusters (37) are detected. The reduced result page size has the added benefit of being accessible from smart phones and tablets (tested for iOS and Android).

antiSMASH 2.0 is currently the most comprehensive software for genome mining and analysis of secondary metabolite biosynthetic pathways, and includes or provides direct links to the most significant other tools and algorithms for this task. The updates to the antiSMASH framework will enable it to be successfully used with the latest sequencing technologies and biochemical insights, while it will continue to be a key tool for state-of-the-art synthetic biology approaches towards secondary metabolism.

Funding

This work was supported by the German Ministry of Education and Research (BMBF) [grant number 0315585A to T.W.]; the German Centre for Infection Research (DZIF) [grant number: 8000-402-2 to T.W.]; and by the Dutch Technology Foundation STW, which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs [grant number STW 10463 to ET]. RB was supported by an NWO-Vidi fellowship.

Supplementary Material

Supplementary tables can be downloaded from http://rdmy.info/ch3

Chapter 4

Detecting sequence homology at the gene cluster level with MultiGeneBlast

Published as:

M.H. Medema, E. Takano, R. Breitling (2013) Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Molecular Biology and Evolution* 30: 1218-1223.

Abstract

The genes encoding many biomolecular systems and pathways are genomically organized in operons or gene clusters. With MultiGeneBlast, we provide a user-friendly and effective tool to perform homology searches with operons or gene clusters as basic units, instead of single genes. The contextualization offered by MultiGeneBlast allows users to get a better understanding of the function, evolutionary history and practical applications of such genomic regions. The tool is fully equipped with applications to generate search databases from GenBank or from the user's own sequence data. Finally, an architecture search mode allows searching for gene clusters with novel configurations, by detecting genomic regions with any user-specified combination of genes.

Sources, pre-compiled binaries and a graphical tutorial of MultiGeneBlast are freely available from http://multigeneblast.sourceforge.net/

Background and Rationale

Many biological systems and pathways from bacteria, archaea and fungi, but also from plants (Field and Osbourn 2008) and animals (Garcia-Fernandez 2005), are encoded by genes that are physically clustered together on the chromosome in operons or gene clusters (Fischbach and Voigt 2010). The architectures of these gene clusters are sometimes well conserved between species, but they may also evolve quickly through rearrangements, insertions, deletions and duplications. In many cases, knowing the evolutionary context of a gene cluster can reveal much about its function, by offering information on which other organisms possess a similar biomolecular system or pathway as encoded by the gene cluster, which parts are most strongly evolutionarily conserved, and what variants of the system or pathway exist. Homology searching can also be useful for mining large numbers of gene or operon variants from homologous gene clusters, which can then function as building blocks for the synthetic biology engineering of novel pathways or systems (Medema et al. 2012).

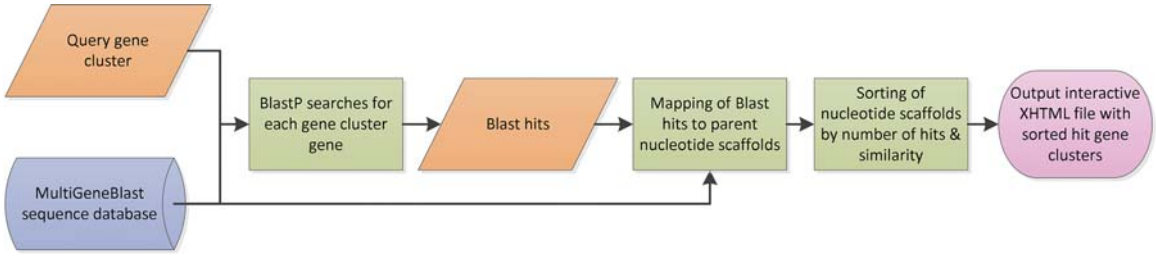

Figure 1: Outline of the homology search process by MultiGeneBlast. First, the amino acid translation of each gene sequence within the query gene cluster is searched against the selected MultiGeneBlast database, yielding a dataset of Blast hits. The Blast hits are then mapped to their parent nucleotide scaffolds, based on the information from the database. Loci detected on the nucleotide scaffolds are then sorted according to their empirical similarity scores with the query gene cluster. Finally, the sorted list of genomic loci is displayed in an interactive XHTML file that can be viewed with any modern web browser.

While several efficient and user-friendly tools are available to perform homology searches for single genes and proteins (e.g., NCBI's BLAST+ implementation (Camacho et al. 2009)), there are few options to exhaustively mine the databases for homologs of entire operons or gene clusters. Tools such as JGI IMG (Mavromatis et al. 2009), PSAT (Fong et al. 2008), CCGV (Revanna, Krishnakumar,

Dong 2009), EDGAR (Blom et al. 2009) and Absynte (Despalins, Marsit, Oberto 2011) each offer the possibility to perform gene neighbourhood comparisons across prokaryotic genomes on precomputed datasets, but none of these allow searches against the entire GenBank database (Benson et al. 2012), nor do they allow generating custom databases from the user's own sequence data. Another tool, SynBlast (Lehmann, Stadler, Prohaska 2008), is restricted to organisms whose genetic information is deposited in ENSEMBL (Flicek et al. 2012).

Here we present MultiGeneBlast, a comprehensive Blast implementation to perform homology searches on multigene modules, which is built as a wrapper around NCBI BLAST+. As with the normal NCBI BLAST+ suite, the user can search the entire GenBank database or create his/her own databases. Additionally, MultiGeneBlast has the ability to perform 'architecture searches', which allow finding genomic loci containing homologs of specific user-specified combinations of genes. Multiple sequence alignments of homologs can be generated automatically after the search, and all results are visualized in a user-friendly interactive XHTML page.

Name	Short description
multigeneblast	Main command-line application to run MultiGeneBlast searches
mgb_gui	Graphical user-interface for configuring and starting a MultiGeneBlast run
makedb	Application to construct MultiGeneBlast databases from user data
makegbdb	Application to construct MultiGeneBlast databases from GenBank divisions
makendb	Application to construct raw nucleotide MultiGeneBlast databases from user data
makegbndb	Application to construct raw nucleotide MultiGeneBlast databases from GenBank divisions
format_embl.py	Script to generate EMBL input files from a genome sequence + gene annotations

Table I: Applications in the MultiGeneBlast package

Implementation of the software

MultiGeneBlast functions as a Python-based wrapper around the blastp program from the NCBI Blast+ suite (Camacho et al. 2009), which allows detecting even distant homology between genes by using the amino acid translation as a proxy for the gene sequence. MultiGeneBlast uses a specific database format in which each FASTA header in the database contains information on the parent nucleotide entry of the protein sequence as well as on the start and end positions and strand orientation of the gene that encodes it — besides, of course, its own functional annotation and accession number. To also make it possible to search unannotated genome sequences for homologous gene clusters, raw nucleotide databases can also be created, on which the tblastn algorithm is used instead of blastp. The MultiGeneBlast implementation (**Figure 1**) extends upon code written earlier for gene cluster comparison in antiSMASH (Medema et al. 2011b).

Setting up a MultiGeneBlast run can be done from the command-line (**Table I**), but also with a user-friendly graphical user interface (**Figure 2**) that allows easy selection of genomic regions (see our

graphical tutorial in **Supplementary File 1**). As in our gene cluster analysis tool antiSMASH, the output is visualized in an interactive XHTML page that can be opened in a web browser. The XHTML page shows an SVG visualization of all sorted genomic loci (**Figure 3**), and clicking on a gene leads to the display of annotation information, details of any blastp/tblastn hit to the (translated) sequence of this gene (percentage identity, sequence coverage, e-value, bit score) and a direct link to run an individual blastp search with the gene encoded by this gene on the NCBI server. Optionally, multiple sequence alignments of the amino acid translations of each query gene sequence with those of its homologs can also be generated using Muscle (Edgar 2004).

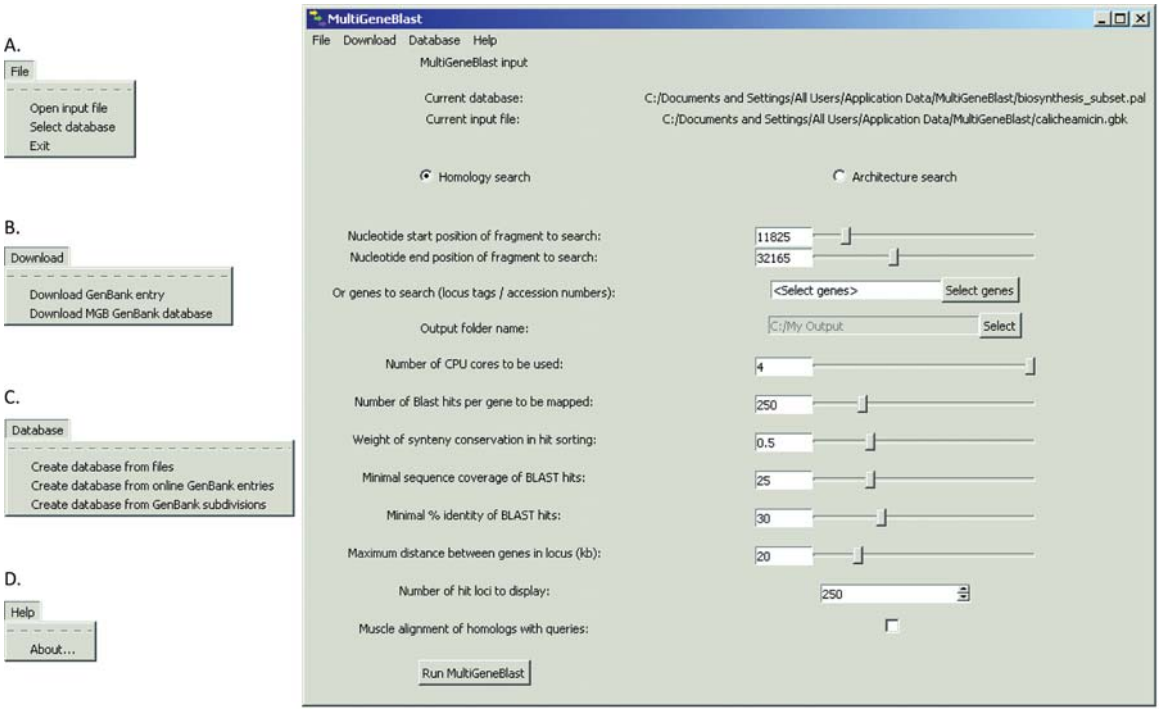

Figure 2: A user-friendly graphical user-interface allows easy construction of databases. A graphical user interface allows easy use of the program. (A) User-friendly selection of input files and databases. (B) Direct download of GenBank entries from NCBI and simple button to download MultiGeneBlast-reformatted GenBank database. (C) Options to design databases from files, from online GenBank entries or from entire GenBank divisions. (D) Link to the MultiGeneBlast website with help pages, a tutorial and various downloads.

Two distinct search modes

MultiGeneBlast offers two distinct search modes: 'homology search' and 'architecture search'. The homology search mode serves to find homologues of a known operon or gene cluster, and hence is an extended version of a standard Blast homology search. The input for a homology search consists of an annotated genome sequence in GBK or EMBL format, together with the start and end locations spanning the query gene cluster or operon. Alternatively to start and end sites, a list of genes can be provided that constitute the gene cluster, which has the advantage that specific genes within the gene cluster can be left out of the analysis. After running separate blastp runs for each amino acid sequence encoded in the query genomic region, MultiGeneBlast locates all hits on their parent nucleotide scaffolds in the database. Each nucleotide scaffold that received blastp/tblastn hits is then subdivided into genomic loci containing blastp/tblastn hits with a maximum mutual distance of

a given number of kb. The default value for this distance is 20kb, a value which has been shown to work well for most bacterial gene clusters (Medema et al. 2011b), but higher values could work better for gene clusters in fungi and plants.

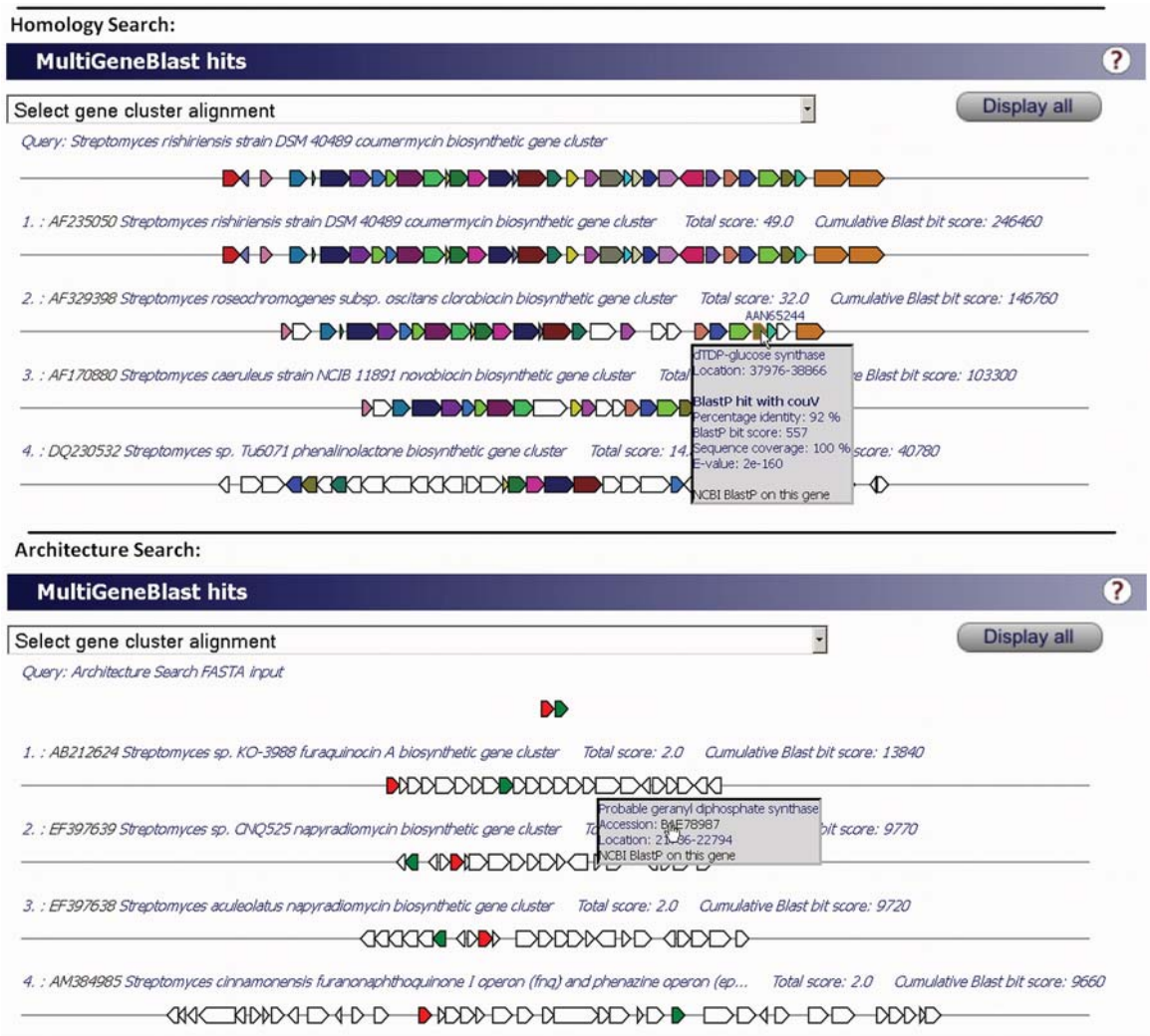

Figure 3: Example output of a MultiGeneBlast run. The output consists of an interactive XHTML page, in which additional information on each gene appears on mouse-over or by clicking on a gene. This feature works for coloured homologous genes as well as white non-homologous genes. The first example output shown here displays a homology search for the coumermycin biosynthetic gene cluster, which identifies gene clusters encoding related compounds. The second example output shows the power of an architecture search to find specific pathways: by using a query of a type III polyketide synthase and a terpene cyclase, biosynthetic gene clusters encoding hybrid polyketide-terpene compounds are identified straightforwardly. Single alignments of the query gene clusters with any particular hit gene cluster can also be selected from a drop-down menu. All gene cluster images are stored in SVG format, so they can easily be transformed in publication-quality figures.

Similar to the clusterblast implementation in antiSMASH (Medema et al. 2011b), genomic loci are then sorted by an empirical similarity score $S = h + i \cdot s$, in which h represents the number of query genes with Blast hits of at least a user-specified sequence coverage and percentage identity to the query, s represents the number of contiguous gene pairs with conserved synteny, and i represents a weighting factor that determines the weight of the synteny in determining the score. The default value for i is 0.5, which gives the number of homologous genes twice the weight as the conservation

of synteny. If the obtained scores are equal, the loci are subsequently sorted by their cumulative blastp/tblastn bit scores. When testing the algorithm on a number of (semi-)manual gene cluster comparisons from the recent scientific literature, we observed that MultiGeneBlast could replicate their results accurately, as well as identify additional homologous but compositionally distinct gene clusters (**Supplementary File 2**).

The architecture search mode differs from a standard homology search in that the query input consists not of a known genomic region, but of a FASTA file with multiple protein sequence entries, designed by the user. Thus, the user can search for all genomic loci containing a combination of certain genes within the same gene cluster. This can be of great use, for example, when searching for gene clusters encoding specific metabolic pathways containing a specified combination of enzymatic steps.

Creating custom databases for MultiGeneBlast

MultiGeneBlast is shipped with a database consisting of the translated amino acid sequences of all gene sequences in the GenBank database (25/07/2012), reformatted with new FASTA headers as stated above. Updated versions of this database will be made available for download regularly. MultiGeneBlast also offers two tools to generate custom databases. The first tool, MakeGBDB, allows the user to construct databases from a specified subset of the GenBank subdivisions (such as BCT for bacteria, PLN for plants, etc.). The tool downloads the specified subdivisions from the NCBI FTP server and then parses them to generate a MultiGeneBlast database. The second tool, MakeDB, allows the user to construct databases from his/her own sequence data, and takes as input a user-specified set of sequence files in GBK or EMBL format. For convenience, a script to generate EMBL files from nucleotide FASTA files and gene annotations is also provided.

New approaches

MultiGeneBlast is the first full-fledged Blast implementation that combines the input of multiple genes into a single query. Compared to previous tools for the comparative analysis of operons and gene clusters, MultiGeneBlast offers a unique set of options (**Table II**).

First of all, MultiGeneBlast, allows to create databases of any combination of published and unpublished data, including the user's personal sequence data. As the costs of DNA sequencing are continuously decreasing, more and more labs have large amounts of unpublished sequence data that need to be analyzed before online publication. No tools have been published thus far that offer the possibility to select the user's own sequence data as both query and subject of the analysis. The Integrated Microbial Genomes (IMG) framework, arguably the most popular tool for gene neighborhood analysis at the moment, by design does not allow any custom queries that are outside the precomputed database, nor does it offer the option to search against custom-designed databases. In contrast, the user-friendly GUI of MultiGeneBlast makes it easy even for biologists with little or no bioinformatic expertise to design their own databases and search them with their own sequence data.

Software	Web tool	Stand-alone tool	Not restricted to precomputed data	Can search entire GenBank database	Based on multiple gene queries	Allows input of personal sequence data	Allows creation of custom databases	Architecture search mode	Command-line available	Open source
MultiGeneBlast		X	X	X	X	X	X	X	X	X
IMG	X									
EDGAR	X									
Absynte	X					X				
PSAT	X									X
CCGV	X		X			X			X	X
SynBlast		X	X						X	X

Table II: Comparison of different software tools for gene cluster homology searches

Secondly, most existing tools do not allow searches against the entire GenBank database but only against subsets of sequences (usually whole genome sequences) for which precomputed results have been obtained. Thousands of known and characterized gene clusters (especially biosynthetic ones) are not part of any whole genome sequence but were instead cloned directly from the environment, or are part of a metagenomic dataset, and are therefore not present in databases such as that of IMG. MultiGeneBlast, however, offers the opportunity to perform a truly exhaustive search to find all homologous genetic elements that are present in the current databases.

Thirdly, the architecture search mode is unique to MultiGeneBlast, and allows finding operons that are not similar to any operon known in advance by the user, but instead contain homologues of a user-specified combination of genes.

Finally, unlike most available tools, MultiGeneBlast can be used from the command-line and also generates a tab-delimited TXT output, so it can easily be integrated into a larger computational pipeline. With relatively simple scripting, large numbers of queries can thus be searched against one or more databases to perform higher-level bioinformatic analyses.

Practical applications of MultiGeneBlast

MultiGeneBlast offers a simple and intuitive tool to perform comparative genomic analysis, facilitating functional inference and evolutionary studies of gene clusters encoding biomolecular machines or pathways.

A major application of MultiGeneBlast is to get a quick overview of the biomolecular diversity of an entire genetic element in diverse organisms, and to survey all the variants that have evolved. Because MultiGeneBlast does not just display the genomic neighborhoods of one single gene but finds genomic loci with a combination of any of a list of query genes, the output will contain variants of the query genomic region consisting of any subset of that region in any arrangement. This avoids

the risk of missing variants that do not contain the query gene, in contrast to approaches based on single gene input. When combining the list of identified gene cluster variants with phylogenetic information (of either species or representative genes), the evolutionary history of a gene cluster can be reconstructed, which can give valuable insight into the biomolecular functions of the various components of the encoded system. Based on patterns of evolutionary conservation, one can sometimes also get a better idea of which genes do and which genes do not belong to the gene cluster as a functional unit.

Often, distinct subclusters with separate evolutionary histories together constitute a larger gene cluster (Fischbach, Walsh, Clardy 2008). A MultiGeneBlast analysis of the entire gene cluster may reveal its fundamental architecture, through the identification of distinct patterns of conservation of various subsets of genes from the gene cluster. This also cannot be achieved by approaches based on a single gene query.

Another important and promising application of the approach is to rapidly harvest gene parts for the synthetic biology design of biochemical pathways (Medema et al. 2011a; Medema et al. 2012). When generating synthetic versions of a particular biochemical system for heterologous implementation in a pre-engineered host, it is of great importance to test multiple versions of the system to find the one that functions best in a particular organism (Bayer et al. 2009). Because MultiGeneBlast can search the entire GenBank database, as well as any personal sequence data that may be available, it can quickly and reliably be used to identify all extant versions of an operon or gene cluster in an exhaustive manner.

Of course, many more applications of the tool are possible, as the co-localization of functionally related genes is a recurring evolutionary motif: MultiGeneBlast provides a general search tool that can be exploited in a wide range of comparative genomics studies of homologous multi-gene units, by expert bioinformaticians and experimental biologists alike.

Acknowledgements

We thank two anonymous reviewers for their constructive comments. We thank Kai Blin for contributing code originally written for the antiSMASH project. This work was supported by the Dutch Technology Foundation STW, which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs (grant number STW 10463). RB is supported by an NWO-Vidi fellowship, and ET by a Rosalind Franklin Fellowship, University of Groningen.

Supplementary Material

Supplementary files can be downloaded from http://rdmy.info/ch4

Chapter 5

Pep2Path: automated matching of peptidogenomics sequence tags to NRPS biosynthetic gene clusters

To be submitted as:

M.H. Medema, Y. Paalvast, P.C. Dorrestein, E. Takano, R. Breitling (2013) Pep2Path: automated matching of peptidogenomics sequence tags to NRPS biosynthetic gene clusters.

Abstract

Nonribosomally synthesized peptides (NRPs) are molecules of great medical importance. They include antibiotics such as penicillin, immunosuppressants such as cyclosporine, and cytostatics such as bleomycin. Recently, an innovative mass-spectrometry-based strategy, peptidogenomics, has been pioneered to rapidly mine microbial strains for novel NRPs using mass spectrometry. Even though the peptide detection can be performed relatively fast, the matching of identified peptide sequences to their biosynthetic gene clusters has remained a bottleneck. Here, we introduce Pep2Path, a simple and efficient algorithm to automate this task. Pep2Path greatly accelerates the peptidogenomics method and thus facilitates a crucial step in the drug discovery pipeline.

Sources and pre-compiled binaries of Pep2Path are freely available from https://sourceforge.net/projects/pep2path/, implemented in Python and supported on MS Windows, Linux and Mac OS X.

Introduction

After a steady decline at the end of the last millennium, natural products are back in the centre of attention as leads for drug discovery (Baltz 2008; Walsh and Fischbach 2010; Winter, Behnken, Hertweck 2011; Zotchev, Sekurova, Katz 2012). The secondary metabolites they are derived from can be categorized into various classes, according to their chemical structures and the different ways in which they are synthesized enzymatically. One of the most abundant classes consists of nonribosomally synthesized peptides (NRPs). Many of these peptides are of great clinical importance, having applications as antibiotics, immunosuppressants or cytostatics (Marinelli 2009).

NRPs are synthesized not by the ribosome but by large enzymatic complexes called nonribosomal peptide synthetases (NRPSs). These NRPSs form assembly lines of modules, which each add one specific amino acid to the growing peptide chain (Fischbach and Walsh 2006). Each NRPS module normally consists of at least three domains: a condensation (C) domain that catalyses the condensation to the next amino acid, the adenylation (A) domain that selects the amino acid to be incorporated, and the thiolation (T) domain to which the growing peptide chain is attached. Because they function independently of the ribosome, NRPSs can introduce not only proteinogenic but also nonproteinogenic amino acids into the peptides they produce. After its synthesis by NRPSs, the core peptide scaffold of an NRP can be further modified by a wide variety of tailoring enzymes.

For a long time, drug discovery from bioactive peptides depended on laborious identification and characterization of one peptide at a time. Moreover, re-discovery of already known peptides occurred with increasing frequency, which made the entire process slow, inefficient and costly (Li and Vederas 2009). Conversely, the genomic identification of biosynthetic gene clusters (BGCs) that could potentially encode the enzymatic machinery to make novel bioactive NRPs is rapidly becoming easier, as thousands of genomes are being sequenced each year and various algorithms have been developed to automatically detect the BGCs that encode NRP biosynthesis (Khaldi et al. 2010; Li et al. 2009; Medema et al. 2011b).

Recently, a new technology, peptidogenomics, has been introduced, which uses the potential of high-throughput mass spectrometry (MS) to speed up the discovery of novel bioactive peptides (Kersten et al. 2011). Using this technology, short amino acid sequence tags (which represent a part of the complete peptide) can be reconstructed from the MS peak patterns by looking at the mass differences between peaks of various peptidic fragments. In turn, these tags can be assessed for their potential to represent novel peptides using dereplication tools such as iSNAP (Ibrahim et al. 2012) and matched to BGCs predicted by methods such as antiSMASH (Medema et al. 2011b). So far, this matching of sequence tags to BGCs has remained a tedious and complicated procedure, in which possible amino acid sequences are manually compared to substrate specificity predictions of NRPSs by algorithms such as NRPSPredictor2 (Röttig et al. 2011), after identification of NRPS gene clusters by antiSMASH. The lack of automation has severely impeded high-throughput peptidogenomic experimentation, and has precluded the use of peptidogenomics on microbial communities with large metagenome datasets. Moreover, the effective use of peptidogenomics on unsequenced strains (Nguyen et al. 2013) also depends on the development of computational approaches, in order to be able to compare identified sequence tags with dozens or even hundreds of genomes of related genome-sequenced strains to identify orthologous BGCs. Here we fill this gap by introducing Pep2Path, a software package that facilitates the rapid and automatic identification of candidate BGCs for amino acid sequence tags detected by mass spectrometry. Moreover, we show how Pep2Path can be used to identify BGCs for previously characterized NRPs for which no biosynthetic mechanism had been elucidated yet.

Algorithm and Implementation

In order to assess how likely it is that a given MS-derived sequence tag originates from a certain BGC, it is necessary to estimate the probability for each amino acid from the tag to originate from each of the NRPS modules from the BGC. Briefly, the derivation of the necessary equations is as follows (see **Supplemental Text** for details): The Bayesian posterior probability $P(M|A)$ that an NRPS module M is responsible for the biosynthesis an observed amino acid A in a peptide sequence tag T is:

$$P(M|A) = \frac{P(A|M) \cdot P(M)}{P(A)} \approx \frac{P(A|M)}{P(A)}$$

For most applications, the prior probability $P(M)$ of a module M to synthesize any observed amino acid will be the same for all modules and can be neglected. The probability $P(A)$ of an amino acid to be present in any position of any sequence tag is estimated using the average frequencies of amino acids in known NRP families within the NORINE database (Caboche et al. 2008), using a pseudocount of 1 to correct for the limited sample size. $P(A|M)$, the probability that a certain amino acid is incorporated by an observed NRPS module M, is calculated as

$$P(A|M) = \frac{P(A) + c \cdot \left({I_{A,M}}^{\eta} + x \cdot P(A) \right)}{1 + c \cdot \left(\sum_{A \in \mathbb{A}} \left({I_{A,M}}^{\eta} \right) + x \right)} \qquad \text{with } x \ll 1$$

in which c is a confidence factor that accounts for how much of the final probability is determined by the substrate specificity predictions, $I_{A,M}$ is a score based on the predicted substrate specificity of the module (see **Supplemental Text** for details), and η allows exponential penalization of

NRPSPredictor2 mismatches. Combining this with the above formula for $P(M|A)$, Pep2Path calculates the score S for a gene cluster C given a sequence tag T as the sum of the log likelihoods $P(M|A)$ for all amino acids in the tag:

$$S(C|T) = \sum_{A\ in\ T} \ln\big(P(M|A)\big) = \sum_{A\ in\ T} \ln\left(\frac{P(A) + c \cdot \left({I_{A,M}}^{\eta} + x \cdot P(A)\right)}{P(A) \cdot \left(1 + c \cdot \left(\sum_{A \in \mathbb{A}} \left({I_{A,M}}^{\eta}\right) + x\right)\right)}\right)$$

Of course, sequence tags can be aligned to a BGC-encoded NRPS assembly line in multiple ways, and BGCs that encode multiple NRPSs may have various assembly-line configurations. Hence, the final score for a match between a BGC and a sequence tag will be the maximum score S obtained for any of the possible alignments of a sequence tag with each of the possible assembly line orders of the NRPSs encoded by a BGC.

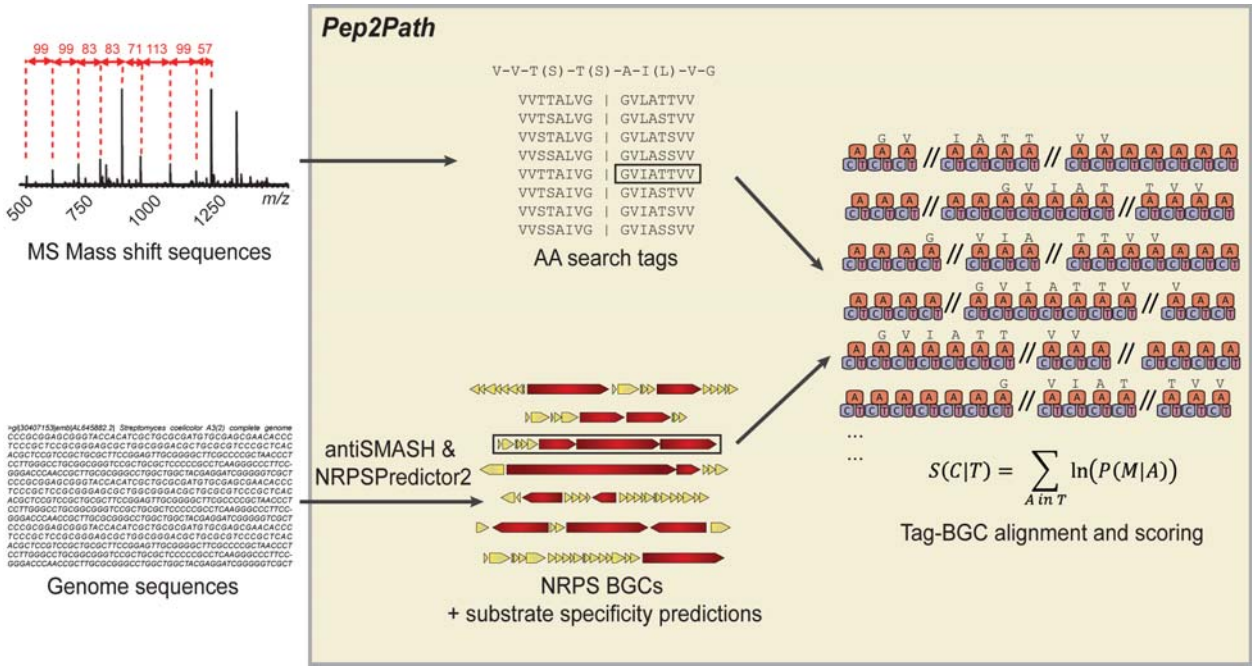

Figure 1: Outline of the Pep2Path matching process. The input for Pep2Path consists of mass shift sequences (or amino acid search tags) on the one hand, and genome sequences on the other hand. The latter are processed into databases by *makedb*, using antiSMASH and NRPSPredictor2. When a database is queried with a mass shift sequence or amino acid search tag, Pep2Path scores all possible matches between search tags and all possible assembly line orders of each of the NRPS BGCs in the database.

The input of genomic data to which sequence tags can be matched proceeds through the construction of Pep2Path databases. For this purpose, two accessory programs, *makedb* and *mergedb*, are shipped with Pep2Path. The *makedb* tool uses antiSMASH2 (Blin et al. 2013) to search user-provided input sequences for BGCs that encode NRPSs and integrates information on these BGCs into a database. Each entry in this database consists of the accession number or name of the nucleotide entry the BGC originates from, a list of genes that constitute the BGC, taxonomic information on the species whose genome encodes it, the modular architecture of the NRPSs within the BGC, and substrate specificity predictions as given by the two NRPSPredictor2 algorithms. A database with all NRPS-containing BGCs within the GenBank database is already provided with Pep2Path. The *mergedb* tool can be used to merge this database with custom-made databases

created from locally available sequence data, or to combine multiple custom-made databases with one another.

The core Pep2Path program uses the formula for $S(C|T)$ to predict links between MS sequence tags and BGCs within a database file. The input can consist of a list of amino acids or of a list of mass shifts. In the latter case, Pep2Path converts the mass shifts into amino acid sequence tags using the conversion table provided by Kersten et al. (2011). Because some amino acids (such as leucine and isoleucine) have identical masses, Pep2path will generate a list of all possible short peptide sequences. Then, Pep2path assesses for each of the BGCs within the selected NRP database how likely it is that this BGC encodes the peptide from which the sequence tag derives (**Figure 1**). Depending on how much is known about the source of the biological material analyzed by MS, the user can select a taxonomic range (strain, species, genus, etc.) within which to search.

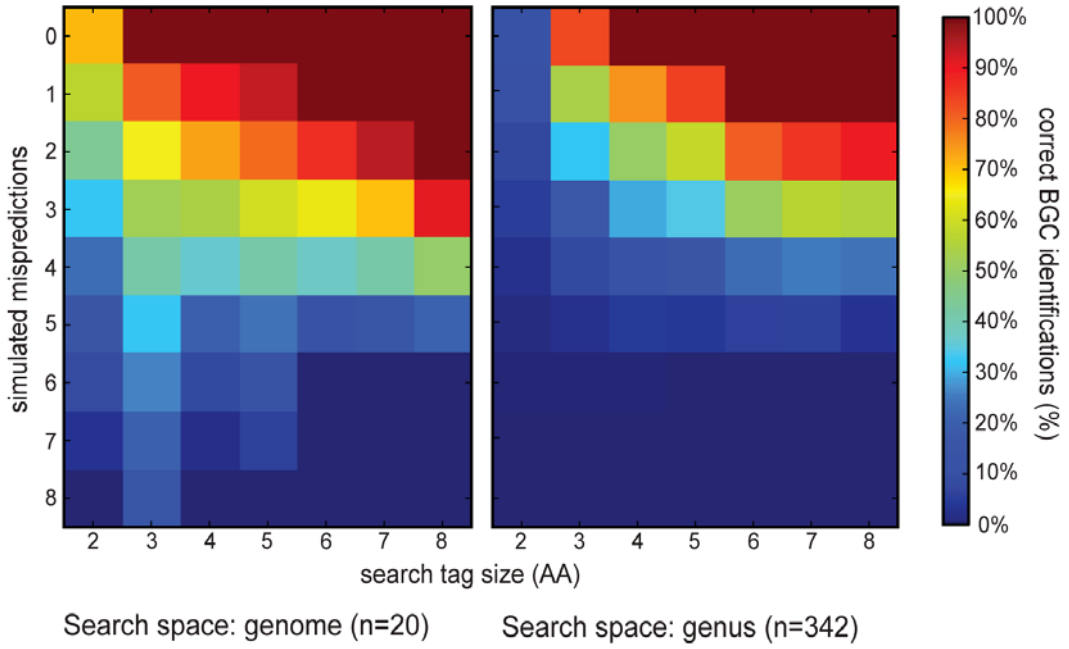

Figure 2: Quality of Pep2Path predictions with varying sequence tag lengths and NRPSPredictor2 prediction qualities. The heat map shows the average number of correct BGC predictions for Pep2Path searches with the stendomycin sequence tag V-V-T(S)-T(S)-A-I(L)-V-G across the *Streptomyces hygroscopicus* ATCC 53653 genome (20 NRPS BGCs) or across all *Streptomyces* nucleotide entries (342 NRPS BGCs). The searches were done for all possible search subtags of 2–8 amino acids long, and for all combinations of 0–8 mispredictions for the corresponding NRPS modules. Mispredictions are simulated with zero scores given by Pep2Path for sequence tags matching to these domains.

Results

As a first performance test, we tested whether Pep2Path would be successful in matching the V-V-T(S)-T(S)-A-I(L)-V-G sequence tag that was used to identify the stendomycin BGC by Kersten et al. (2011). Pep2Path appeared to do this very effectively: even when the search space consisted of all 4357 NRPS BGCs in the GenBank database, the stendomycin BGC emerged as the top hit. Of course, the stendomycin case presents a relatively favorable situation, in which the NRPSPredictor2 predictions are good and the sequence tag is long. To investigate how well tag-BGC matching would work with smaller sequence tags or with NRPSPredictor2 mispredictions, we varied the sequence tag

size between 2 and 8 and introduced 0–8 mispredicted substrate specificities for the corresponding stendomycin NRPS modules, simulated by artificial zero scores given to sequence tag amino acids that are matched to these modules. The results show that BGCs with good NRPSPredictor2 predictions can be identified robustly within a genome or species even with small sequence tags (**Figure 2**). Only for BGCs with more than 25–50% mispredicted substrate specificities, Pep2Path will often fail.

tag size (AA)	BGC search space size: 5	BGC search space size: 10	BGC search space size: 25	BGC search space size: 50	BGC search space size: 100
2 (*n*=18)	75%	64%	47%	36%	26%
3 (*n*=15)	78%	70%	54%	44%	37%
4 (*n*=12)	83%	78%	65%	56%	45%
5 (*n*=11)	90%	89%	79%	72%	61%
6 (*n*=8)	96%	96%	87%	81%	74%
7 (*n*=5)	99%	99%	96%	91%	88%
8 (*n*=3)	100%	100%	100%	100%	100%

Table I: Benchmark of Pep2Path on 18 recently discovered NRPS BGCs. For each tag size, all possible search tags of that size in the test set of peptides (**Supplementary Table I**) were used as queries. For each BGC search space size, 50 search spaces were generated from randomly selected BGCs from the same (sub)phylum that the NRP originates from. The resulting percentages represent the average number of cases in which the correct BGC ended up as the (shared) best hit across all possible sequence tags and across all possible search space permutations. Shared best hits were included because of the frequent presence of orthologous BGCs encoding the same molecule in related genomes. The *n* in the left column signifies the number of test peptides large enough to be included in the analysis for this tag size.

To investigate how well Pep2Path would work in practice on novel compounds, we mined 18 NRPs from the recent scientific literature, together with their corresponding NRPS BGCs (**Supplementary Table I**). None of these had yet been incorporated into the NRPSPredictor2 training sets. In order to assess how well Pep2Path would be able to match tags from these NRPs to the correct BGCs under varying conditions, we varied the sequence tag size from 2 to 8 and tested all possible search tags of these sizes on databases with sizes ranging from 5 NRPS BGCs to 100 NRPS BGCs (an average bacterial genome contains ~5 NRPS BGCs). For each database size, 50 randomly permuted BGC databases were created from BGCs originating from genomes within the same (sub)phylum, and the results were averaged across all of these permutations. The results (**Table I**) confirm that the minimum sequence tag size to confidently match an NRP to a BGC is around 2–4 when the genome sequence is known. When the search space is larger, a situation that represents the mining of unsequenced strains for NRPs and attempting to match them to orthologous BGCs within the same genus (Nguyen et al. 2013), larger sequence tags (e.g., 5–8 amino acids long) are often still sufficient to identify the correct BGC.

Finally, we used Pep2Path in an effort to find BGCs for NRPs within the NORINE database (Caboche et al. 2008) for which no BGC had been discovered so far. Intriguingly, we discovered novel candidate BGCs for five molecules by searching all NRPS BGCs from the species from which the compound had originally been isolated (**Table II**). When we expanded the search space to screen the entire database, we discovered another very good match, for tripropeptin A (**Table II**). Although this eight-amino-acids-long peptide was originally discovered in the gamma-proteobacterium *Lysobacter* sp. (Hashizume et al. 2001), we found a match with a BGC in the genome of the beta-proteobacterium *Collimonas fungivorans* Ter33. This BGC had eight NRPS modules, of which seven gave NRPSPredictor2 predictions exactly matching the tripropeptin A sequence in the right order, while the eighth prediction was only a near miss (ornithine predicted instead of arginine). It is highly probable that this gene cluster encodes the biosynthesis of a tripropeptin, which suggests that the

gene cluster has undergone horizontal gene transfer at least once, from one subphylum to another. All in all, the results of the NORINE searches show how the use of Pep2Path can generate new candidates for experimental testing, even in the absence of new peptidogenomic data.

Compound	Reference	Species (accession nr.)	Locus tags	NRP search tag from NORINE	NRPSPredictor2 prediction	Pep2Path score (rank)
azotobactin delta	(Demange et al. 1988)	*Azotobacter vinelandii* DJ (NC_012560)	Avin_25560-Avin_25650	(ChromophA)-asp-ser-hse-gly-asp-ser-cit-hse-orn-hse	(glu-nrp-nrp)-asp-ser-nrp-gly-asp-ser-arg-nrp-orn-nrp	8.99 (1)
trichotoxin	(Irmscher et al. 1978)	*Trichoderma virens* Gv29-8 (ABDF02000085)	TRIVIDRAFT_69940	ala-gly-ala-leu-ala-glu-ala-ala-ala-ala-ala-ala-pro-leu-ala-xxx-gln-vol	nrp-nrp-ala-nrp-nrp-gln-nrp-ala-nrp-ser-leu-nrp-pro-nrp-ala-ala-gln-vol	6.25 (1)
ferintoic acid	(Williams et al. 1996)	*Mycrocystis aeruginosa* 9701 (CAIQ01000336)	MICAK_4000004-MICAK_4000007	trp-co-lys-val-hty-ala-phe	phe-nrp-lys-val-nrp-ala	5.24 (1)
plusbacin	(Shoji et al. 1992)	*Pseudomonas putida* ND6 (CP003588)	YSA_0461-YSA_0481	asp-pro-ser-asp-arg-pro-ala-allothr	asp-ser-ser-asp-nrp-nrp-nrp-thr	4.91 (1)
amphibactin B	(Martinez et al. 2003)	*Vibrio tubiashii* NCIMB 1337 (AHHF01000067)	VT1337_12727-VT1337_12732	orn-orn-ser-orn	orn-orn-ser-orn	2.73 (1)
tripropeptin A	(Hashizume et al. 2001)	*Collimonas fungivorans* Ter33 (NC_015856) Originally found in *Lysobacter* sp.	CFU_2182-CFU_2185	thr-pro-pro-arg-asp-ser-pro-asp	thr-pro-pro-orn-asp-ser-pro-asp	8.94 (1)

Table II: Novel matches of NORINE-derived NRPs to BGCs detected in genome sequences. Candidate BGCs for azotobactin delta, trichotoxin, ferintoic acid, plusbacin and amphibactin B were discovered by searching within the taxonomic range of the species in which the molecules were found. The candidate BGC for tripropeptin A was discovered by searching the entire Pep2Path database.

Conclusion

With Pep2Path, we introduce an automated approach for a key technology to accelerate natural products research. An automated peptidogenomics technology constitutes a radical departure from the one-molecule-per-experiment approach to drug discovery from natural products that has dominated the field for long. The combination of the rapid sampling of chemical space by tandem mass spectrometry with Pep2Path's computational effectiveness now makes it possible to simply isolate a strain, extract the cellular content, put it in a mass spectrometer and sequence the DNA, and then click a button on the computer to detect a collection of novel peptides along with their associated BGCs. Even more intriguingly, one could leave out the DNA sequencing and culture extraction at first, and simply match MS spectra from environmental samples directly to (cryptic) gene clusters in the database of already sequenced genomes (Nguyen et al., in press). Strain isolation and (partial) genome sequencing could then be performed only in the cases when the peptide seems promisingly novel.

When applied to large-scale metagenomic datasets, the combined use of mass spectrometry and Pep2Path will also make it possible to sample NRPs in environmental samples at large scales, by integrating mass spectral molecular networking (Guthals et al. 2012; Watrous et al. 2012) from MS data with BGC similarity networks (Cimermancic, Medema et al., in preparation), and cross-linking these two networks with a third type of edges based on Pep2Path matching scores between molecule families and gene cluster families.

As the approach can potentially be extended to other compound classes such as RiPPs (ribosomally synthesized and post-translationally modified peptides) and saccharides, Pep2Path paves the way

for the high-throughput characterization of the vast universe of BGCs discovered in genomic data of organisms throughout the tree of life.

Acknowledgements

This work was supported by the Dutch Technology Foundation STW, which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs [STW 10463]. RB is supported by an NWO-Vidi fellowship.

Supplementary Material

Supplementary text can be downloaded from http://rdmy.info/ch5

Chapter 6

MultiMetEval: comparative and multi-objective analysis of genome-scale metabolic models

Published as:

P. Zakrzewski*, M.H. Medema*, A. Gevorgyan, A.M. Kierzek, R. Breitling, E. Takano (2013) MultiMetEval: comparative and multi-objective analysis of genome-scale metabolic models. *PLoS ONE* 7: e51511.

*Equal contribution

Abstract

Comparative metabolic modelling is emerging as a novel field, supported by the development of reliable and standardized approaches for constructing genome-scale metabolic models in high throughput. New software solutions are needed to allow efficient comparative analysis of multiple models in the context of multiple cellular objectives.

Here, we present the user-friendly software framework Multi-Metabolic Evaluator (MultiMetEval), built upon SurreyFBA, which allows the user to compose collections of metabolic models that together can be subjected to flux balance analysis. Additionally, MultiMetEval implements functionalities for multi-objective analysis by calculating the Pareto front between two cellular objectives. Using a previously generated dataset of 38 actinobacterial genome-scale metabolic models, we show how these approaches can lead to exciting novel insights. Firstly, after incorporating several pathways for the biosynthesis of natural products into each of these models, comparative flux balance analysis predicted that species like *Streptomyces* that harbour the highest diversity of secondary metabolite biosynthetic gene clusters in their genomes do not necessarily have the metabolic network topology most suitable for compound overproduction. Secondly, multi-objective analysis of biomass production and natural product biosynthesis in these actinobacteria shows that the well-studied occurrence of discrete metabolic switches during the change of cellular objectives is inherent to their metabolic network architecture.

Comparative and multi-objective modelling can lead to insights that cannot be obtained by normal flux balance analyses. MultiMetEval provides a powerful platform that makes these analyses straightforward for biologists.

Sources and binaries of MultiMetEval are freely available from https://github.com/PiotrZakrzewski/MetEval/downloads.

Introduction

Living cells owe their existence to complex metabolic networks, in which large numbers of chemical conversions occur to allow the cells to harvest energy, sustain themselves and reproduce. In the past decades, methodologies have been developed to systematically describe and quantitatively analyse (parts of) the metabolic network of a cell in computational models (Covert et al. 2001; Edwards, Ibarra, Palsson 2001). Such reconstructions have already been of great use to develop a better understanding of the metabolic architecture and dynamics of various organisms (Breitling, Vitkup, Barrett 2008; Oberhardt, Palsson, Papin 2009).

Genome-scale constraint-based metabolic models are reconstructions of metabolism that comprise the stoichiometries of all reactions predicted from whole genome sequences based on the presence of enzyme-coding genes. Accordingly, they can be used to model the steady-state behaviour of the metabolism of a whole organism (Feist et al. 2009; Price, Reed, Palsson 2004). Well-accepted procedures on how to generate genome-scale constraint-based models are available, based on Enzyme Classification annotations and generic gap-filling procedures (Thiele and Palsson 2010).

The resulting metabolic models can be used to perform several kinds of analyses (Lewis, Nagarajan, Palsson 2012; Price et al. 2003), the most popular one being flux balance analysis (FBA; Orth, Thiele, Palsson 2010). In this method, the fluxes of metabolites through the network are calculated based on the stoichiometry of each reaction and an objective function that specifies for which goal (e.g., maximization of biomass production from a given input or minimization of nutrient uptake) the fluxes are optimized.

Recently, high-throughput methods have been developed to generate and gap-fill metabolic models for multiple species in a rapid and standardized way (Henry et al. 2010), based on genome annotations obtained with a uniform method. Even though the resulting models still need to be compared with experimental data to achieve optimal quality (Kim et al. 2012) and the gap-filling implemented by SEED is not always optimal (Brooks et al. 2012; Latendresse et al. 2012), automatically generated models that have undergone a limited amount of manual curation are already useful for obtaining a rough assessment of the metabolic capabilities of cellular systems.

The standardization offered by automation opens up the road for comparative modelling, as little model reconciliation is needed, in contrast to what is usually the case for manually reconstructed models (Oberhardt et al. 2011). Comparative analysis of genome-scale metabolic models is an intriguing new field with diverse potential applications (Alam et al. 2011; Blank, Lehmbeck, Sauer 2005). For example, it can be used to detect evolutionary differences between metabolic networks of related species and predict their relative adaptive ecological value (Mithani, Hein, Preston 2011). It can also be used to assess the suitability of a range of species for a particular biotechnological application (e.g., biofuel or drug production) based on the topologies of their metabolic networks, which could then inform the choice of industrial production hosts (Lee, Pandu Rangaiah, Lee 2010).

As well as studying multiple models at the same time, it can also be very revealing to optimize models for multiple objectives simultaneously (Nagrath et al. 2010; Oberhardt et al. 2010). Many different 'natural' objective functions have been proposed, such as the maximization of biomass, secondary metabolite production or ATP production, or the minimization of total flux, redox potential or nutrient uptake (Feist and Palsson 2010). For most of these, there are reasons to believe that the cellular flux distribution can be expected to have evolved in a way that optimizes the objective, at least under specific conditions. It can even be argued that evolution has driven biological systems toward an optimal compromise between all of these, sometimes conflicting, objectives. Other relevant objective functions that one would like to consider are those that correspond to the aims of bioengineering instead of evolution, such as the maximization of the production of a specific metabolite. Unfortunately, as implementing different objective functions is relatively difficult in most existing analysis platforms, many published studies have been restricted to exploring a single objective function. Usually, this objective function is the maximization of biomass production, although interesting studies have been performed that explore different objective functions (Feist et al. 2010). Some of these have been made available through the COBRA toolbox (Schellenberger et al. 2011)).

A pair of objective functions (such as a biomass objective function and the objective function of production of a specific compound) can be balanced to find the so-called Pareto front (Marler and Arora 2004) between the two objectives. The Pareto front comprises the set of "Pareto-optimal" solutions, for which one objective can only be improved at the expense of the other objective.

Bacterial metabolism has recently been shown to operate close to such Pareto fronts (Schuetz et al. 2012). An analysis of such a front enables one to predict the interactions between different metabolic processes and priorities within the cell. For example, one can identify the extent to which two objectives compete for the use of the same enzymatic pathways. Moreover, one can use the results to predict the balance between the objectives that is optimal for sustaining biomass levels while producing as much of a certain valuable metabolite as possible.

Here, we describe a new software package, Multi-Metabolic Evaluator (MultiMetEval), a simple framework that provides an efficient and user-friendly interface for the comparative study of multiple models and the use of multiple objective functions. The software has been conveniently linked up to the SurreyFBA package for metabolic modelling (Gevorgyan et al. 2011), allowing for easy interaction with general modelling algorithms. In order to make the tool widely useful, it includes a new global SBML Level 2 parser that enables input of models from popular modelling platforms, including SEED (DeJongh et al. 2007; Henry et al. 2010), KGML (Kanehisa et al. 2004) and COBRA (Becker et al. 2007), overcoming previous compatibility issues between different SBML flavours that severely impaired comparative analyses. Moreover, all functionalities are organized in a graphical user-interface that allows the user to quickly generate publication-quality plots from the results and export the results for downstream analyses in other software packages.

In a case study, we show how the principles of comparative modelling can be applied to a concrete biological problem with our software, in a comparative study of the metabolic networks of 38 actinobacteria. Based on the 38 genome-scale models, we predict the suitability of different bacterial strains for the heterologous production of a range of different secondary metabolites and use multi-objective analysis to study the dynamic balance between the biomass objective and the compound production objective. We find that the maximally attainable fluxes to a natural product vary greatly between species as well as between the chemical classes of compounds. Moreover, we observe discrete switch-like behaviour in the models when the priority of the compound production objective function is gradually increased compared to the biomass objective function; this provides a possible systems-level explanation for the metabolic switches observed in the onset of secondary metabolism in such organisms (Alam et al. 2010).

Design and Implementation

The MultiMetEval comparative analysis framework was written in Java 6 Standard Edition with an interface handled by the Swing framework and integrated plot generation handled by the JFreeChart library. It is functional on both Windows and Linux operating systems. The program was built upon the SurreyFBA framework (Gevorgyan et al. 2011), which is used as an engine for the basic FBA calculations. Additionally, in order to read input models from a large range of sources (e.g. SEED (DeJongh et al. 2007; Henry et al. 2010) and COBRA (Becker et al. 2007)), a Python-based universal SBML parser was generated to convert input SBML files into a valid SurreyFBA input format. Combined with the parser and the SurreyFBA engine, MultiMetEval allows for high-throughput comparative and multi-objective analysis of metabolic models that share the same syntax.

Parsing of input SBML files

Incompatibility of SBML models coming from different frameworks has been a major drawback for comparative studies (Oberhardt et al. 2011). SBML Level 2 itself is a general-purpose language for systems biology, and can be used for storing a great number of data types. There is, however, still no universally adopted definition of FBA-specific parameters within the SBML namespace. Therefore, gene-protein-reaction association rules and reaction capacity bounds have to be defined using annotations and general parameters. This leads to many different format varieties of SBML, in which the data relevant for FBA are stored in different ways.

Existing FBA frameworks make use of their own parsers enforcing usage of their own SBML format variety. In order to make SBML files from different frameworks cross-compatible in our tool, we generated a parser that can convert any major SBML format variety into the SurreyFBA format. As we show in **Table I**, our parser adds a flexibility that has not been possible in the other major FBA tools. In principle, our parser could easily be implemented in other contexts as well.

Framework	**SEED-generated SBML**	**KGML-derived SBML**	**COBRA-generated SBML**
MultiMetEval	+	+	+
COBRA	–	+/–	+
VANTED	–	+	–
SurreyFBA 1.0	–	+/–	+

Table I. Comparison of parsing capabilities of MultiMetEval with other FBA frameworks. Table showing SBML parsing abilities of the most popular FBA tools. Only the MultiMetEval parser is able to successfully process SBML models from SEED (DeJongh et al. 2007; Henry et al. 2010), KGML (Kanehisa et al. 2004) and COBRA (Becker et al. 2007).

Comparative analysis

MultiMetEval provides a user-friendly facility to perform comparative analysis of multiple metabolic models, by combining batch runs of the single model analysis functionalities provided by SurreyFBA with new features that allow for convenient multi-model input and output.

The basic units analysed by the comparative analysis module are “model collections”, which are sets of models selected by the user for analysis and parsed into the same format by the universal SBML parser.

A specific menu allowing user-friendly construction of such collections is available via the File menu. To allow reuse of models in different collections, the collections can be created as subsets of a main model repository that holds all models that were imported to the program. Models can easily be added to the main repository and then moved to any collection in the same window.

For every model, the number of reactions, metabolites, orphan reactions and orphan metabolites are detected and displayed in MultiMetEval’s main overview table when a collection is opened. FBA can be performed on the entire collection at once by clicking a simple menu button, and results are output in a single spreadsheet table (**Figure 1**).

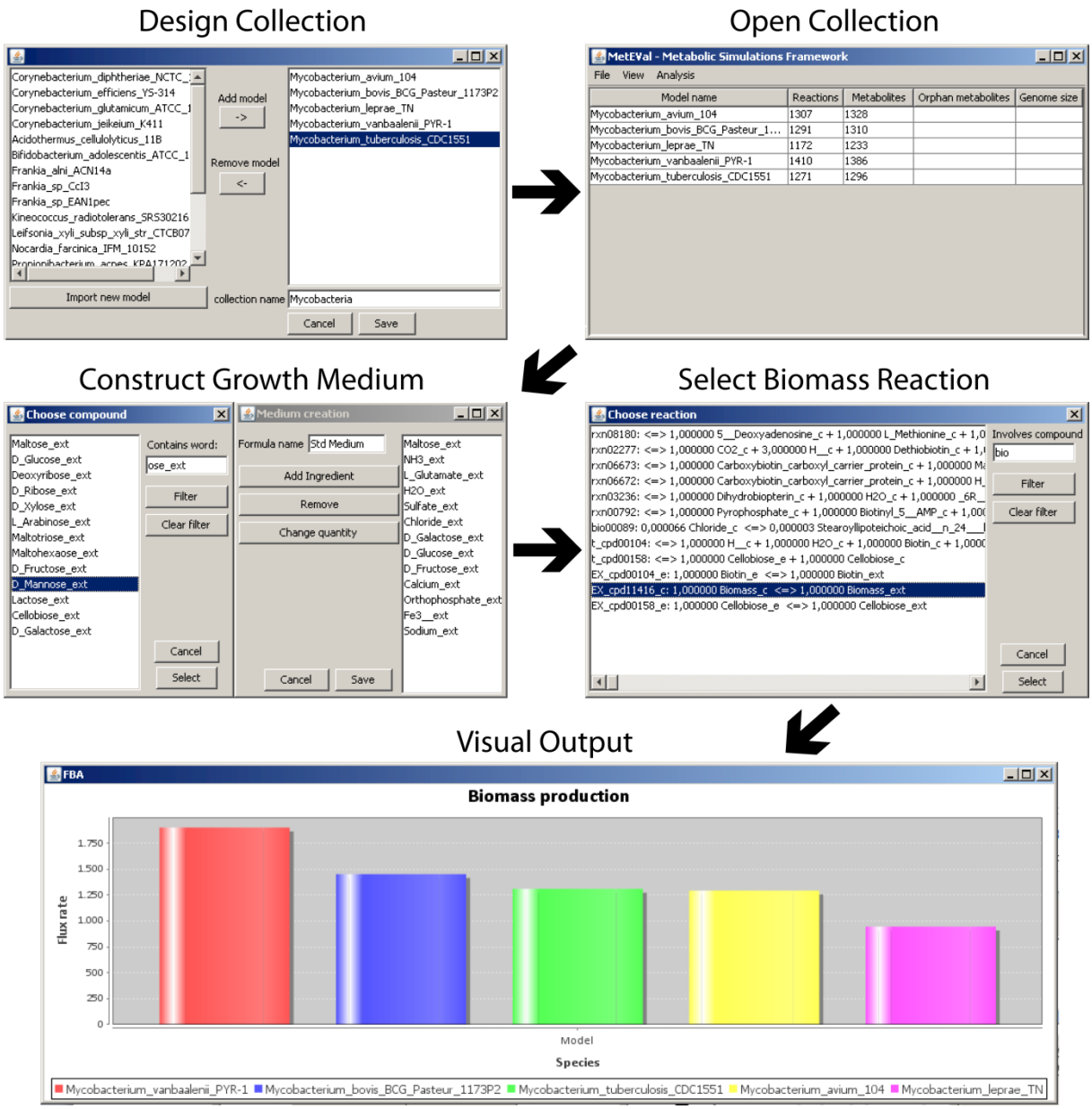

Figure 1. Workflow of comparative metabolic analysis in MultiMetEval. First, a collection of metabolic models is designed and a growth medium is constructed. Then, flux balance can be performed on all models in the collection simultaneously.

Our framework also offers a growth medium editor, which allows comparative analysis not only of different models, but also in different growth conditions. In order to make sure that the medium is compatible with the models, the medium description format used in our framework operates only on the metabolites present in a given collection and restricts the choice of medium ingredients to those which were defined in any of its models as external. The motivation for this is, of course, that these are the only metabolites that can be consumed by at least one model.

Multi-objective analysis by Pareto front calculation

MultiMetEval allows performing multi-objective analysis by calculating the Pareto front (Oberhardt et al. 2010; Vo, Greenberg, Palsson 2004) for maximization of two given reactions. Compared to the

weighted sum approach (which was already implemented in SurreyFBA), Pareto front calculation is more informative, as it avoids the arbitrary nature of weight assignment.

In this analysis, MultiMetEval calculates the tradeoff between two objectives. Often, the first objective will be the biomass production reaction, but, in principle, MultiMetEval can calculate a Pareto front for the optimization of any combination of two fluxes of reactions that co-occur in the same model.

In the Pareto front calculation, first the maximal possible flux of the first objective is calculated. This value will be used in the following steps as a constraint that is iteratively decreased at each step. So after calculating the maximal flux of objective one, the program will carry out optimizations for the second objective *n* times (were *n* represents the resolution), and with each simulation step the constraint put on the reaction by the first objective will decrease unless its value reaches zero.

The results of the multi-objective analysis are output to a results table as well as in a visual plot (**Figure 2**).

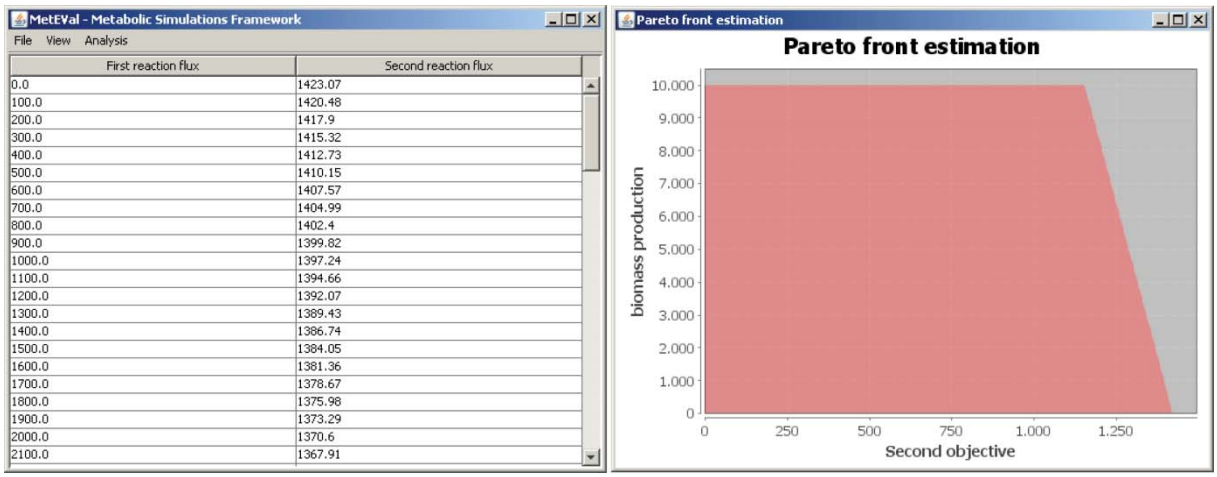

Figure 2. Table and plot output from the Pareto front calculation routine. The first steps are identical to those in Figure 1, except that only one organism is selected and two reactions are selected to calculate their trade-off.

In addition to the implementation of the Pareto trade-off routine in the MultiMetEval framework described here, we also implemented it in the SurreyFBA command-line interface as well as in JyMet, the single-model analysis framework from SurreyFBA (Gevorgyan et al. 2011), for additional flexibility.

Results/Discussion

Comparative and multi-objective metabolic modelling has many exciting applications in systems and synthetic biology (Durot, Bourguignon, Schachter 2009; Medema et al. 2012; Oberhardt, Palsson, Papin 2009). To illustrate the power of these approaches, we applied the MultiMetEval tools in an exemplary case study on the production of secondary metabolites in actinobacteria. We show how comparative FBA can be used to identify differences between organisms in their theoretical production capacities for such metabolites, as well as differences in the extent to which biomass production competes with secondary metabolite biosynthesis.

Comparative FBA of secondary metabolite biosynthesis by 38 actinobacteria

In our comparative FBA analysis, we constructed a model collection in MultiMetEval from the 38 genome-scale metabolic models of actinobacteria that were recently constructed and curated by Alam et al. (2011) (excluding the two *Tropheryma* models, but including models for *Bifidobacterium adolescentis* ATCC 15703, *Bifidobacterium longum* NCC2705 and *Kineococcus radiotolerans* SRS30216). We then reconstructed biosynthetic pathways for 15 secondary metabolites of different classes that were present as annotated pathways in the KEGG database. These included polyketides (erythromycin, tylosin, aureomycin, tetracycline), aminoglycosides (butirosin, neomycin, streptomycin), aminocoumarins (clorobiocin, coumermycin, novobiocin), nonribosomal peptides (enterobactin, pyochelin, cephalosporin, penicillin) and a beta-lactam (clavulanic acid). Such types of compounds are highly relevant biotechnologically, because they often have antimicrobial or anti-cancer activities (Fischbach and Walsh 2009). Their biosynthetic pathways can be (re-)engineered with synthetic biology approaches and expressed for purposes of drug discovery and industrial production (Medema et al. 2011a). For each of the 15 metabolites, derivative models were then made for all 38 actinobacteria, in which the biosynthetic pathway for the metabolite was added to the genome-scale model. For all $38 \cdot 15 = 570$ models, FBA was then performed using MultiMetEval on a minimal medium with equal amounts of glucose as the sole carbon source, ammonium as the sole nitrogen source, and orthophosphate as the sole phosphorus source. The cellular objective was maximization of the production of the secondary metabolite. A limited number of (maximally seven) reactions for glucose uptake and methionine biosynthesis, as well as compound-specific reactions for precursor biosynthesis were added to each model to enable it to produce the compound on the minimal medium (see **Supplementary Table I**).

Figure 3 shows the resulting heat map representing the theoretical maximal production rates of the 15 secondary metabolite classes in all 38 actinobacteria. The intensity of a colour depicts the relative flux rate – the lighter the colour (closer to white), the higher a given flux value is in comparison to others from the same column.

As no regulatory and kinetic information is used in the constraint-based models, one should note that the variation observed between the species only represents the difference due to differences in network topology given the medium composition used. Still, it is intriguing that substantial differences in theoretical production capacities are observed between the actinobacterial species. As expected, we observe some correlation with general topological properties of the metabolic networks such as the numbers of reactions and metabolites: minimalistic genomes generally tend to be less efficient predicted production hosts (e.g., *Bifidobacterium* and *Propionibacterium*). However, these differences clearly do not account for all the variation observed. Among the most interesting exceptions is the severely genome-minimized *Mycobacterium leprae*, which still reaches surprisingly high predicted fluxes. Members of the same class of secondary metabolite (which also have similar precursors) are usually predicted to be most efficiently produced in the same hosts. An exception is formed by two nonribosomal peptides, cephalosporin and enterobactin, for which fewer species are able to obtain the maximum observed flux towards compound production than for two other nonribosomal peptides, pyochelin and penicillin. This is probably due to the requirement of additional precursors, 2-oxoglutarate and 2,3-dihydroxybenzoic acid, respectively, for these two molecules, which are not required for penicillin and pyochelin.

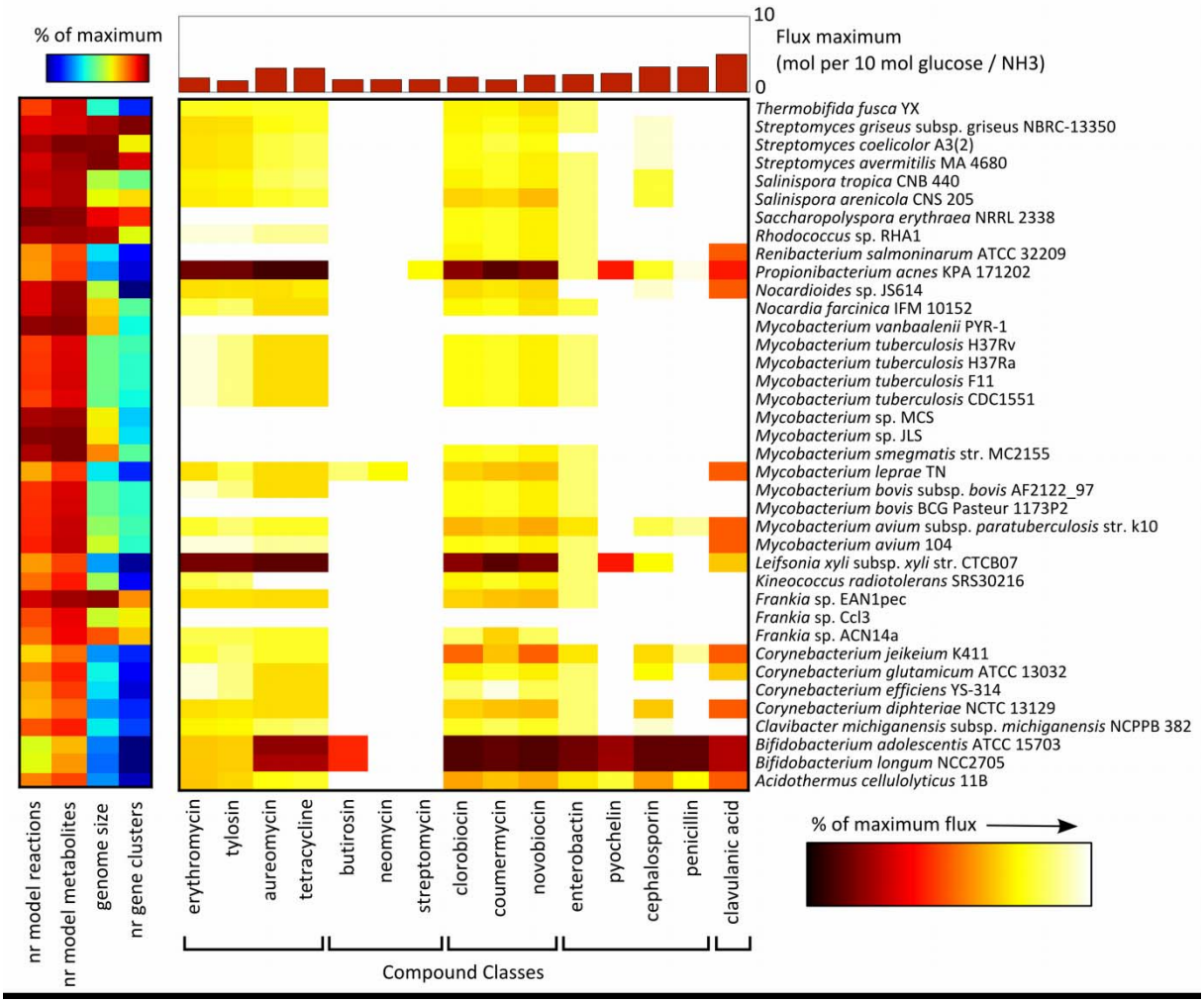

Figure 3. **Theoretical maximum fluxes of secondary metabolite production.** The heat map shows relative maximal fluxes of the final biosynthetic step in the metabolic pathways leading 15 different secondary metabolites, which were incorporated into the genome-scale metabolic models of 41 actinobacteria. Flux balance analysis was performed on the minimal medium described by Alam et al. (2011). White indicates a high relative flux level, red indicates a low relative flux level (as % of the maximally obtained value across all species, displayed at the top of the figure). In the heatmap on the left, the number of model reactions and metabolites, the genome sizes and the number of secondary metabolite biosynthesis gene clusters (predicted using antiSMASH, Medema et al. 2011b) are plotted.

When we investigated the presence of which reactions influences fluxes most, by calculating the correlation between reaction presence/absence and maximum fluxes for each compound (**Supplementary Table II**), we could observe that in at least a number of cases this corroborated current biochemical knowledge. For example, clavulanic acid fluxes most strongly correlate with the presence of a reaction (rxn00101) to convert urea into CO_2 and NH_3, which corroborates the unusual presence of a microbial urea cycle in its native host organism *Streptomyces clavuligerus* (Ives and Bushell 1997; Romero, Liras, Martin 1986). Also, the fluxes towards several compounds (the macrolides, aminocoumarins and pyochelin) correlated with the presence of a reaction (rxn00141) converting S-adenosylhomocysteine to adenosine and homocysteine, which corroborates evidence for a positive effect of S-adenosylmethionine regeneration on antibiotic biosynthesis (Zhao, Gust, Heide 2010).

Interestingly, the fact that the genome of a species has a lot of secondary metabolite biosynthetic gene clusters does not necessarily mean that its metabolic network is optimized for a higher

production of such metabolites compared to other species: high metabolite diversity does not imply high metabolite production titers. The fact that streptomycetes, such as *Streptomyces coelicolor* and *Streptomyces avermitilis*, famous for their production of a wide variety of clinically and biotechnologically important secondary metabolites (Bentley et al. 2002; Ikeda et al. 2003) and endowed with about 25-30 gene clusters per genome, do not score particularly highly suggests that their metabolic networks may not have been optimized for achieving high production titers of such metabolites during their evolution. This may partly explain why extensive metabolic engineering and classical strain optimization have usually been essential to optimize production strains for economically viable metabolite production, often with tremendous improvements in titres (Adrio and Demain 2006; Medema et al. 2011c; Yanai, Murakami, Bibb 2006).

On the other hand, models representing species from the taxonomic branch of free-living mycobacteria (*Mycobacterium vanbaalenii*, *Mycobacterium* sp. MCS, and *Mycobacterium* sp. JLS) achieve the highest predicted production rates for secondary metabolites in the simulations, although they have only about 15 secondary metabolite gene clusters per genome. The difference with the pathogenic *Mycobacterium* species, such as *M. tuberculosis*, *M. bovis* and *M. leprae*, may be explained by the further genome minimization of the pathogenic species, which may have led to a loss of flexibility in the metabolic networks.

Generally, comparative modelling as described here could lead to a more systematic approach towards the identification of suitable “universal hosts” for heterologous expression of gene clusters (Alexander et al. 2010; Freitag et al. 2006; Stevens et al. 2010). Specifically, this preliminary analysis already suggests that free-living mycobacteria might be an attractive starting point for the generation of a minimal actinobacterial genome for use in synthetic biology approaches (Pfeifer and Khosla 2001; Scherr and Nguyen 2009), especially as all three of them belong to the fast-growing mycobacteria.

As expected, similar patterns of theoretical maximal production rates across organisms were observed for compounds with similar chemical structures, such as the aminocoumarins novobiocin, coumermycin and clorobiocin. Also notable is that the metabolic networks of some organisms appear more fit for the production of certain compounds than others. For example, *Renibacterium salmoninarum* ATCC 32209 is predicted to be one of the best producers of polyketides and one of the worst producers of clavulanic acid. This suggests that the species differences observed are not caused by the presence or absence of single enzymes, but that different factors play a role for different compound types. Some aspects that could play a role are 1) the presence or absence of pathways directed towards the necessary precursors (metabolic detours are probably energetically costly), 2) efficiency of ATP generation from the used carbon source glucose, and 3) the ability of models to re-utilize the (sometimes quite exotic) side products of biosynthetic pathways to generate more precursors.

Of course, it should be kept in mind that this study used only mildly curated automatically generated metabolic network models to illustrate the main concepts of comparative flux balance analysis, and a more careful manual curation will be needed before committing substantial experimental resources to testing the hypotheses suggested here. Additionally, more systematic analysis of the specific differences between topologies associated with high and low production capacities of the different compound types may offer specific leads for metabolic engineering, by revealing

topological bottlenecks. Another interesting follow-up study would consist of designing several additional media to study the dynamic interactions between network topology and medium composition or environmental niche.

Analysis of the trade-off between secondary metabolite biosynthesis and biomass production

Biotechnological optimization of natural product biosynthesis often suffers from pathway competition with fluxes leading to the synthesis of biomass components (Gonzalez-Lergier, Broadbelt, Hatzimanikatis 2005; Paradise et al. 2008). In order to assess competition between secondary metabolite biosynthesis and biomass production for selected key species and metabolites, we used multi-objective analysis to calculate Pareto fronts between the biomass objective and the compound production objective.

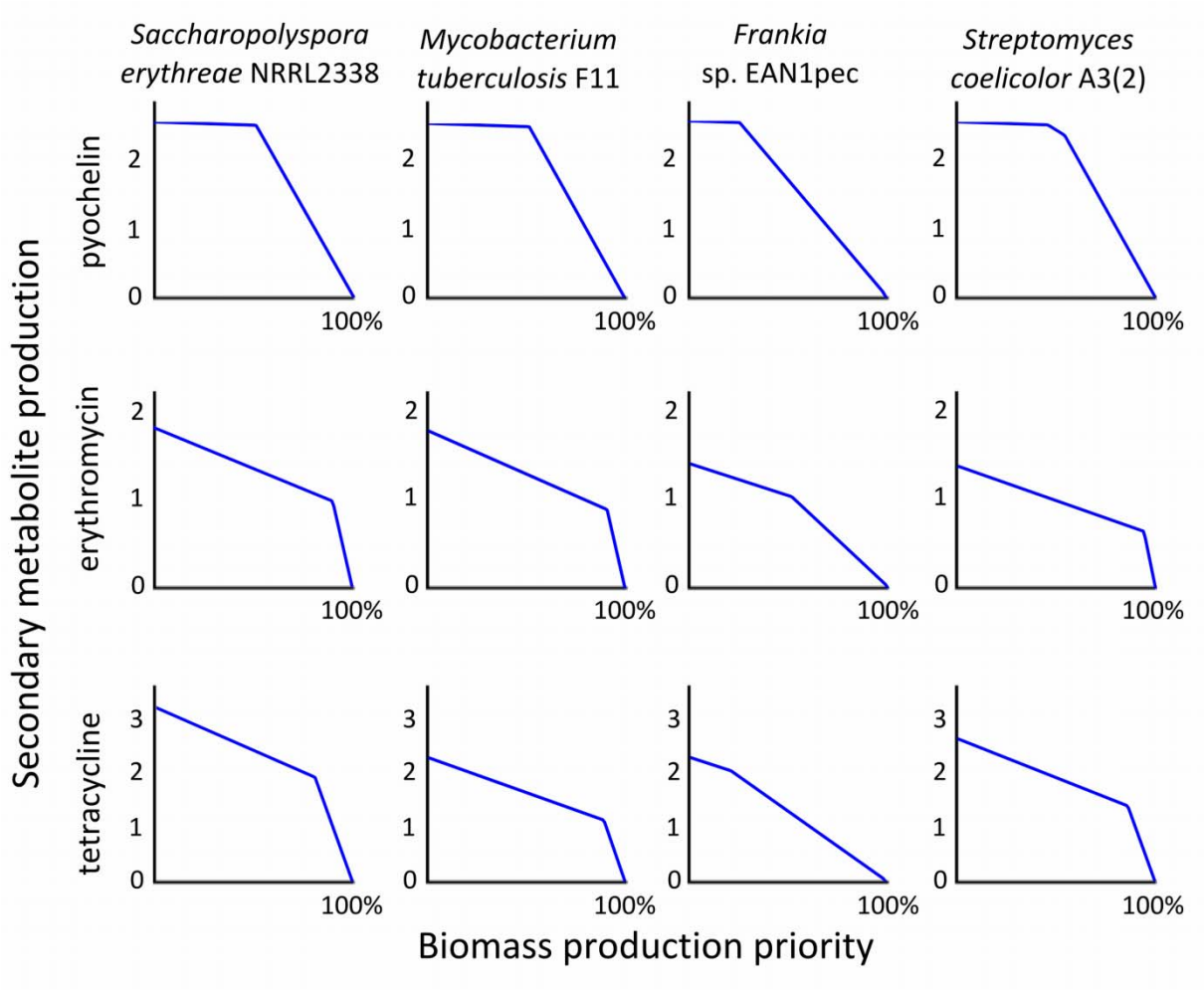

Figure 4. **Pareto front calculation between biomass production and secondary metabolite biosynthesis.** Pareto fronts are given for four species and three different natural products. To estimate secondary metabolite production, the flux rate through the final step in the biosynthetic pathway of the corresponding compound was used as a proxy.

In **Figure 4**, the y-axis on each plot represents the flux rate through the final biosynthesis reaction in the pathway for production of the given compound. The x-axis on each plot represents the

percentage of maximal biomass production flux achieved. The region underneath the blue line represents the space of feasible solutions.

The predictions suggest that the overlap between network resources needed for biomass production and compound production differs notably between species, even without taking into account the organism-specific biomass compositions. It is likely that this has to do with the rate with which the network topology of a species enforces pathway competition between the two objectives, and to which extent alternative pathways are available for both processes.

The applied multi-objective analysis thus characterizes organism-specific relationships between biomass production and compound biosynthesis. Methods such as OptKnock (Burgard, Pharkya, Maranas 2003) can subsequently enable metabolic engineers to reach a position close to the identified Pareto front, by determining how the compound production objective can be optimized given a certain biomass rate by for example stoichiometrically forcing the strain to synthesize a target compound as a by-product of growth.

In the simulations for pyochelin, biomass production at biomass maintenance levels (the almost horizontal plateau at the beginning of the curves) hardly competes with compound production. It appears that in this case the production of biomass components from the medium leaves several metabolic resources unused at the point where the first nutrient limitation from the medium prevents higher biomass production. We confirmed this by recalculating the trade-off under several different medium conditions. Indeed, we observed that pyochelin production and biomass production were constrained by different nutrient limitations: orthophosphate and NH_3 were the limiting medium ingredients, respectively. When medium influx bounds of these compounds were increased by 100% each, the horizontal plateau disappeared.

In that sense, there is a “free lunch” for compound production as long as it is limited by a different nutrient than biomass production is. Remarkably, this suggests that production titers of industrial strains can sometimes be optimized without costs to the biomass maintenance.

In most plots, a single transition point is observed, at which the production titre starts to drop much more drastically when biomass production is increased. This might signify that the metabolic networks of these microbes have at least two distinguishable states in which a different nutrient is limiting for compound production given the fixed biomass production flux at that point. A “metabolic switch” seems to operate at this point, at which the regulation of metabolism probably needs to be drastically changed to maintain optimal levels of both biomass and compound production (i.e., to remain near the Pareto front). Of course, switch-like behaviour would be expected given that FBA is based on linear programming, and different linear constraints will be limiting at different points in the graph. Nonetheless, the fact that the switches corroborate observations from experimental microbiology, in which a carefully regulated switch has been observed at the onset of secondary metabolite biosynthesis (Alam et al. 2010; Nieselt et al. 2010), suggests that cells may employ regulatory mechanisms to remain very close to such a theoretical polygon-shaped Pareto front (Schuetz et al. 2012).

Conclusions

Comparative metabolic modelling is a new field, and as with any recent advance in biology, new software solutions are needed to achieve its full potential. With MultiMetEval, we provide an easy-to-use software framework to analyse large collections of metabolic models in parallel and to perform multi-objective analysis, coupled to the SurreyFBA framework. Although this is just a starting point for further software development, the tool already allowed us to study secondary metabolism in actinobacteria in novel ways. Most interestingly, comparative analysis of their genome-scale models predicts that the organisms whose genomes encode the largest numbers of biosynthetic gene clusters do not necessarily have the metabolic network topology most suited for industrial production of these compounds, suggesting an interesting line of enquiry for future experimental work. Additionally, results from multi-objective analysis suggest that bacterial metabolic switches are not just enforced by regulation, but are grounded in the very architecture of the metabolic system in which they occur. We expect that further experimental analysis will likely give exciting definitive insights into these phenomena.

Acknowledgements

We thank Tauqeer Alam for providing the actinobacterial genome-scale metabolic models.

Supplementary Material

Supplementary tables can be downloaded from http://rdmy.info/ch6

Chapter 7

Computational tools for the synthetic design of biochemical pathways

Published as:

M.H. Medema, R. van Raaphorst, R. Breitling, E. Takano (2012) Computational tools for the synthetic design of biochemical pathways. *Nature Reviews Microbiology* 10: 191-202.

Abstract

As the field of synthetic biology is developing, the prospects for *de novo* design of biosynthetic pathways are becoming more and more realistic. Hence, there is an increasing need for computational tools that can support these efforts. A range of algorithms has been developed which can be used to identify all possible metabolic pathways and their corresponding enzymatic parts. These can then be ranked according to various properties and modelled in an organism-specific context. Finally, design software can aid the biologist in the integration of a selected pathway into smartly regulated transcriptional units. Here, we review key existing tools and offer suggestions for how informatics can help shape the future of synthetic microbiology.

Introduction

A key promise of synthetic biology is the possibility to customize the metabolic system of microorganisms for the commercial production of a wide diversity of high-value biofuels (Bond-Watts, Bellerose, Chang 2011; Steen et al. 2010) or natural products (Khalil and Collins 2010; Medema et al. 2011a; Walsh and Fischbach 2010). Pathways for the production of alcohols, biodiesels, polyketides and terpenoids have successfully been constructed by introducing combinations of parts from various origins into a bacterial host that is easy to cultivate (Ajikumar et al. 2010; Atsumi, Hanai, Liao 2008; Hanai, Atsumi, Liao 2007; Menzella et al. 2006; Steen et al. 2010; Zhang, Li, Tang 2008). Potentially, entire metabolic pathways can be (re-)designed *in silico* and implemented in specialized host organisms (Martin et al. 2009; Prather and Martin 2008; Soh and Hatzimanikatis 2010; Tyo, Kocharin, Nielsen 2010; Weeks and Chang 2011). Successes obtained in pioneering work on the anti-malarial drug artemisinin (Chang et al. 2007; Martin et al. 2003; Ro et al. 2006) suggest that such approaches can be very fruitful. A biosynthetic pathway towards this compound was successfully engineered in *Saccharomyces cerevisiae* and *Escherichia coli* (see **Box 1**), which has the potential to enable much more cost-effective production of this important drug than before, when it could only be harvested in a laborious and costly fashion from the source plant *Artemisia annua*.

The experimental work involved in engineering a synthetic pathway takes a considerable amount of effort, and even systematically planned experiments are usually accompanied by much trial and error. When conceiving the design of a novel biosynthetic pathway (**Figure 1**), the synthetic biologist has to find optimal solutions in selecting pathways, enzymes or host organisms from an abundance of possibilities. In this review, we will explore how the use of powerful computational tools (**Table I**) can lead to better-informed and more rapid design and implementation of novel pathways, and we will propose ways in which tools from different fields of computation can be linked together effectively. We will discuss the different methodologies to identify all possible metabolic pathways that can lead to the synthesis of a compound of choice, and how to rank these pathways based on various criteria. Subsequently, we will consider how flux balance analysis of pathways can be applied to identify the most suitable candidate host organisms. We will also examine how to effectively search the sequence databases to obtain a list of candidate parts (genes, operons, etc.) for the execution of each step in the proposed pathway. Finally, we will discuss how computational

methods can aid in refactoring these parts and integrating them into well-designed transcriptional units optimized for a specific host organism.

For specific case studies and more detailed explanations on the inner workings of each of the computational methods, we refer to a range of excellent specialist reviews that have been published recently (Feist et al. 2009; Marchisio and Stelling 2009; Soh and Hatzimanikatis 2010; Xu 2009).

System	Description
Pathway Prediction	
BNICE (Hatzimanikatis et al. 2005)	Biochemical Network Integrated Computational Explorer; framework for identification and thermodynamic assessment of all possible pathways for the degradation or production of a given compound.
System of Cho et al. (2010)	Framework for identification and prioritization of biosynthetic pathways for the synthesis of a user-specified chemical.
DESHARKY (Rodrigo et al. 2008)	Pathway identification algorithm, which identifies pathways that best match up to the native metabolic network of a specific host, and provides the user with amino acid sequences of corresponding enzymes from phylogenetically closely related organisms.
RetroPath (Carbonell et al. 2011)	Web server hosting a unified framework for retrosynthetic pathway design, integrating pathway prediction and ranking, prediction of compatibility with host genes, toxicity prediction and metabolic modeling.
FMM (Chou et al. 2009)	From Metabolite to Metabolite; web server that finds biosynthetic routes between two metabolites within the KEGG database.
CarbonSearch (Heath, Bennett, Kavraki 2010)	Algorithm that identifies pathways within existing metabolic networks by tracking the conservation of atoms moving through them.
OptStrain (Pharkya, Burgard, Maranas 2004)	Computational framework that advises on optimization of the host's metabolic network to add a particular metabolic pathway, by adding or deleting reactions.
Parts Identification	
Registry of Standard Biological Parts	MIT parts registry, containing various types of biological parts such as promoters, ribosomal binding sites, transcriptional terminators, and plasmids; the registry mostly contains parts collected during iGEM competitions.
Standard Biological Parts knowledgebase (Galdzicki et al. 2011)	Knowledgebase with parts (including all parts from the Registry of Standard Biological Parts), which have been transformed into Synthetic Biology Open Language to make the information computable.
IMG (Mavromatis et al. 2009)	Integrated Microbial Genomes; environment for the comparative and evolutionary analysis of microbial genomes, including gene neighbourhood orthology searches.
antiSMASH (Medema et al. 2011b)	Identification, annotation and comparative analysis of secondary metabolite biosynthesis gene clusters.
KEGG (Kanehisa et al. 2002)	Key collection of databases of metabolites and metabolic pathways; includes organism-specific and general maps of metabolic pathways and networks, gene-enzyme associations, orthology information, and more.
ASC (Röttig, Rausch, Kohlbacher 2010)	Active Site Classification; uses a protein structure to find residues near the active site of enzymes, which it uses to construct suppor-vector machines that classify subclasses (e.g., substrate specificities) of enzymes within an enzyme family.
Parts Refactoring and Synthesis	
RBS Calculator (Salis 2011)	Automated design of ribosome binding sites based on a thermodynamic model of transcription initiation.
RBSDesigner (Na and Lee 2010)	Algorithm for prediction of mRNA translation efficiencies, as well as design of ribosome binding sites for a desired protein expression level.
Gene Designer 2 (Villalobos et al. 2006)	Software package for gene, operon and vector design, codon optimization and primer design
GeneDesign (Richardson et al. 2010)	Web server with algorithms ranging from codon optimization and codon bias graphing to insertion of restriction sites into a protein-coding nucleotide sequence and designing building blocks based on restriction site overlaps.
Gene Composer	Commercial software suite for genetic construct design, codon optimization and gene assembly
OPTIMIZER (Puigbo et al. 2007)	Web server that performs codon optimization on an input protein-coding DNA sequence, using a codon usage table.
DNAWorks (Hoover and Lubkowski 2002)	Web server for oligo design for PCR-based gene synthesis, with integrated codon optimization.
TmPrime (Bode et al. 2009)	Web server for oligo design for PCR-based gene synthesis, with integrated codon optimization.
CloneQC (Lee et al. 2010)	Web application for quality control on sequenced clones, by detection of errors in DNA synthesis.
Pathway and Circuit Design Software Packages	

Biojade	Software tool for design and simulation of genetic circuits
Clotho	Flexible interface for synthetic biological systems design; within the interface, a range of apps/plugins can be utilized to import, view, edit and share DNA parts and system designs.
Tinkercell (Chandran, Bergmann, Sauro 2009)	Computer-aided design software, which allows drag-and-drop drawing and simulation of biological systems.
Asmparts (Rodrigo, Carrera, Jaramillo 2007)	Computational tool that generates models of biological systems by assembling models of parts
GenoCAD (Czar, Cai, Peccoud 2009)	Computer-aided design software for design of multigene DNA sequences, with the option to assist the user through interactive 'grammar checking' of the design drafts.
WebGEC	Web simulator from Microsoft for genetic circuit design and testing
SynBioSS (Weeding, Houle, Kaznessis 2010)	Software suite for designing, modeling and simulating synthetic genetic constructs; the SynBioSS Designer can be used to transform a sequence of BioBricks (from the Registry of Standard Biological Parts) or other parts to model that can be simulated in the SynBioSS Desktop Simulator.
CellDesigner	Editor for graphical drawing regulatory and biochemical networks, which can be stored in Systems Biology Markup Language (SBML)
BioNetCAD	CellDesigner plug-in for computer-aided design and simulation of biochemical networks
Metabolic Modeling / Flux Balance Analysis	
COBRA Toolbox (Becker et al. 2007)	Standard toolbox for metabolic modeling and FBA
SurreyFBA (Gevorgyan et al. 2011)	Command-line tool and GUI for constraint-based modeling of genome-scale networks
CycSim (Le Fevre et al. 2009)	Web server for analyzing genome-scale metabolic models; includes enzyme knockout simulations.
BioMet Toolbox (Cvijovic et al. 2010)	Web toolbox for analyzing genome-scale metabolic models; includes gene knockout analysis, flux optimization, and more.
iPATH 2 (Yamada et al. 2011)	Interactive visualization of data on metabolic pathways; items on KEGG-based metabolic maps can be coloured based on the user's preferences.
GLAMM (Bates, Chivian, Arkin 2011)	Interactive visualization of data on metabolic pathways, can use host-specific metabolic networks and allows detection of pathways within a network.

Table I. Key computational tools currently available for pathway construction

Prediction and prioritization of possible pathways

For compounds of biotechnological value, often only a single specific biosynthetic pathway has been characterized. The key promise of the synthetic biology approach to pathway design is, however, that one does not remain limited to biosynthetic routes that already exist in nature. Instead, realistic biosynthetic pathways can be constructed from first principles to optimize their thermodynamic efficiency.

During the last decade, a range of computational pathway prediction algorithms have been generated which can aid in pathway (re-)design. Some predictors focus on changing existing pathways through making knock-outs or adding novel enzymes (Pharkya, Burgard, Maranas 2004). Other predictors have been built to identify possible metabolic pathways from first principles (Hou, Wackett, Ellis 2003; McShan, Rao, Shah 2003), based on possible biotransformations between chemical structures. More recently, several algorithms have been constructed that use more complex search heuristics to find and rank all possible pathways leading to a desired end compound (**Table I, Figure 1**).

Software for metabolic pathway identification and ranking

One accessible and user-friendly system for pathway identification is "From Metabolite to Metabolite" (FMM), a freely available web service where one can search possible pathways between

known input and output compounds (Chou et al. 2009). It combines the KEGG maps and KEGG LIGAND information to form combined pathway maps, identifies the corresponding genes and organisms and gives an output in which different pathways can be compared. A drawback of this system is that it is limited to characterized pathways present within the (often incomplete) KEGG framework, and that it does not give further insight into the practical or thermodynamic feasibility of the pathway. However, the fact that it can quickly give a clear overview of different possible metabolic routes towards a product of interest can make it a convenient starting point for many investigations.

Box 1: Experimental successes in synthetic pathway engineering

Several pioneering experimental efforts in the construction of synthetic pathways have highlighted the potential of the field.

Arguably the most famous example is that of artemisinin, a potent anti-malaria drug that is naturally produced by the plant *Artemisia annua*. As large-scale production of the compound from plant biomass is very difficult, synthetic biologists instigated a project to engineer its biosynthetic pathway in the bacterium *E. coli*. Already in 2003, researchers succeeded in introducing a yeast-derived pathway for the production of isoprenoids artemisinin precursors in *E. coli* (Martin et al. 2003). Later, they also succeeded in developing a synthetic pathway consisting of plant- and microbe-derived enzymes that was capable of producing artemisinic acid (which can be converted into artemisinin in just two chemical steps) at high titers in *E. coli* and *S. cerevisiae* (Chang et al. 2007; Dietrich et al. 2009; Ro et al. 2006). In a largely similar fashion, others have successfully introduced a plant-derived pathway to produce taxadiene, the first committed intermediate for the anticancer drug taxol, in *E. coli* (Ajikumar et al. 2010). After carefully balancing the expression of the heterologous pathway and the native pathway producing the necessary isoprenoid precursors, production levels were increased more than 10,000-fold.

Another elegant early example of synthetic engineering of biosynthetic pathways is displayed in the work of Müller et al. (2006), who engineered a pathway for the biosynthesis of D-hydroxyphenylglycine, an important building block for the side chain of semi-synthetic penicillins and cephalosporins. They combined a hydroxymandelate synthase and a hydroxymandelate oxidase from *Streptomyces coelicolor* and *Amycolatopsis orientalis* with a stereo-inverting hydroxyphenylglycine aminotransferase from *Pseudomonas putida*. Although the yields that were obtained initially were not very high, the results highlighted the potential of combining enzymes from various biological sources into a novel pathway.

Regarding biofuel production, synthetic pathways for re-routing bacterial native metabolism towards the production of isopropanol and higher alcohols were introduced into *E. coli* in a similar fashion, by testing enzymes from a range of different organisms (including engineered versions of native enzymes) and finally expressing the combination that had been tested to result in the highest yields (Atsumi, Hanai, Liao 2008; Hanai, Atsumi, Liao 2007). More elaborate synthetic approaches, which also entailed the redesign of specific transcriptional units and simple regulatory circuits in combination with introducing enzymes from other microbes, later lead to the production of biodiesels and waxes in *E. coli* directly from simple sugars (Steen et al. 2010).

Perhaps even more intriguingly, Bayer et al. (2009) engineered an efficient pathway for the synthesis of methyl halides in a full-fledged synthetic manner. They selected all 89 putative homologues of the enzyme methyl halide transferase from bacteria, plants, fungi and archaea that were identified by a Blast search on the entire NCBI sequence database. Subsequently, they designed codon-optimized versions of all of them using Gene Designer. Finally, they used these to chemically synthesize a synthetic gene library that could be tested to find the enzyme that performed the desired function most effectively in the host strain, which resulted in production titers up to 190 mg/lh.

Dunlop et al. (2011) engineered microbial biofuel export and tolerance by creating a similar synthetic library of hydrophobe/amphiphile efflux transporters. In addition to a simple Blast homology search, they used substrate specificity predictions based on the specificity-determining regions of the transporters to generate a subset of 43 homologues that represented a uniform distribution across all candidates. In this study, the genes in the synthetic library were not codon-optimized, and this could be the reason that the native *E. coli* gene still ranked highest in a large number of assays performed. An additional codon optimization step could have led to even more impressive results.

A more advanced method, BNICE (Hatzimanikatis et al. 2005), predicts novel pathways based on the somewhat broader reaction rules of the Enzyme Commission (EC) classification system. Because BNICE is not restricted to entries from a specific database, it can also predict unknown pathways that are potentially chemically feasible. In its search for pathways, it takes into account the starting compound and/or product, the requested length of the pathway and the range of reactions searched over. The last criterion means that one can choose to only search for a pathway using enzyme reactions out of one known pathway, a combination of multiple pathways or the whole

metabolic network. This makes it easy to perform a more targeted search for shorter, more efficient pathways. BNICE can be a good first step in finding possible pathways, but a lot of subsequent analysis of the results is needed to obtain a useful outcome. This is a conscious choice of the developers: because the BNICE framework is restricted to the first analysis steps, it can be applied to various different kinds of investigations including not only engineering of novel pathways, but also metabolic pathway analysis and pathway retrosynthesis (Bachmann 2010).

In some searches, BNICE predicts more than 10,000 different pathways for the biosynthesis or degradation of a certain compound, due to the few criteria the system relies on. Therefore, it is of paramount importance to not only predict possible pathways, but also rank them based on discriminative criteria. Recently, Henry et al. (2010) have pioneered a prioritization approach within the BNICE framework, by ranking novel 3-hydroxypropanoate biosynthesis pathways by thermodynamic feasibility, pathway length, maximum achievable yield, and maximum achievable activity. In this manner, they could obtain an informative ranking of the otherwise randomly ordered list of thousands of predicted pathways. Interestingly, the currently commercially used pathway was among the top 4 pathways in the ranked list, but it was matched (and in some aspects exceeded) by three novel pathways that could provide interesting alternative designs for industrial implementation.

Another prediction system which is based on enzymatic reactions, but which approaches the search for novel pathways quite differently, is DESHARKY (Rodrigo et al. 2008). The major difference concerns the choice of host organism for the pathway, which for DESHARKY is the first step of the pathway prediction after compound design. The algorithm searches for all possible pathways that connect the metabolic network of the organism to a target compound, after which the thermodynamic favourability and the energy loss in transcription and translation are calculated. The tool is most useful if the host organism of choice has already been determined and one needs to search for the pathway towards a certain compound that will work most efficiently in that organism.

Cho et al. (2010) have constructed a unified system that both predicts and ranks pathways. Starting with a database of reactions categorized by type and a database of reaction rules describing different reactions, the system first predicts a wide range of possible pathways. A ranking algorithm then prioritizes the pathways based on binding site covalence (similarity of reactions in terms of chemical structure changes), chemical similarity, thermodynamic favourability, pathway distance and organism specificity.

However, rankings only address the symptoms, not the cause of the explosion of possibilities — devising a method to flexibly constrain the search space would be a major breakthrough. The recently published web server RetroPath (Carbonell et al. 2011) offers such a principled way to manually determine the strictness of the initial search for reactions. It searches based on molecular signatures of the compounds and reactions involved. Each signature has different ‘heights’, which correspond to levels of structural detail. By varying the height, one can retrieve numbers of reactions varying from the large numbers of reactions resulting from BNICE to the small number of original reactions that are present in the KEGG database. RetroPath also hosts several interesting additional features for filtering and ranking, such as predictions of promiscuous activity of enzymes, predictions of the compatibility between host and heterologous genes, and compound toxicity predictions.

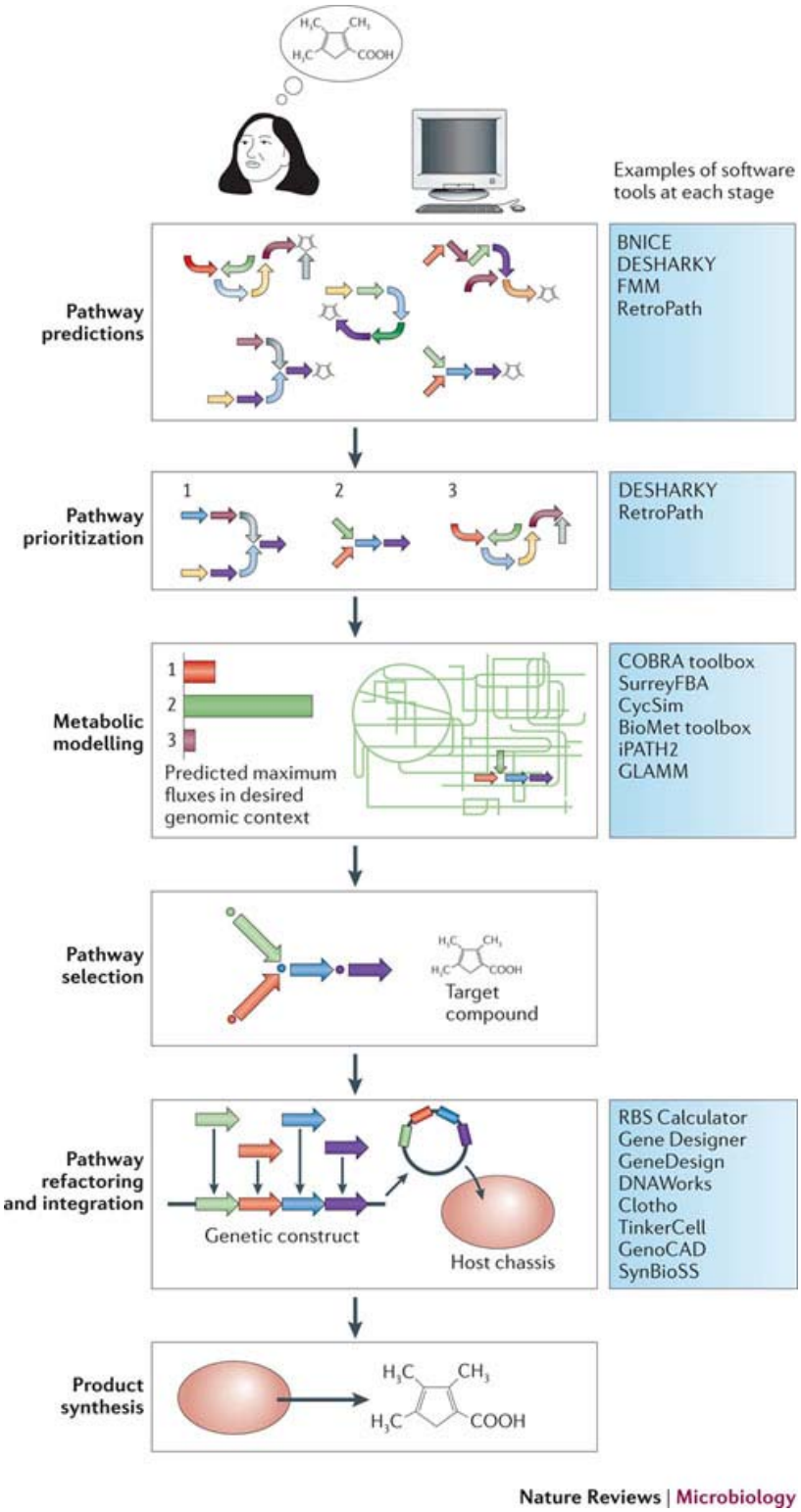

Figure 1. Generalized workflow for *de novo* engineering of biosynthetic pathways, from initial idea to final product. First, a large number of possible pathways is predicted based on chemical reaction rules and/or metabolic maps. Subsequently, the resulting pathways are prioritized based on a number of criteria. Comparative modelling is then performed to predict theoretical production capacities of candidate host organisms for the compounds using the different pathways in the context of the topology of their metabolic network. Finally, one or a few suitable pathways are selected for which synthetic expression constructs can be designed. A diversity of computational tools is essential for all steps, as indicated schematically at the right side of the figure. The order of the steps is not necessarily linear; various iterations and feedback

loops between the steps may be necessary for optimization, as sometimes information obtained in a 'later' stage suggests revision of an 'earlier' decision.

Once a pathway has been selected for introduction into a specific host bacterium, the consequences of this manipulation have to be predicted in the new metabolic context. One system, which aims to monitor the effects of the new pathway on the host, is OptStrain (Pharkya, Burgard, Maranas 2004) – a method that uses flux analysis to give advice on how production could be optimized by altering the host's gene expression. After constructing a hypothetical biosynthetic pathway towards the target compound, the OptStrain system changes the pathway in such a way that as many enzymes from the pathway as possible are native to the host organism. With the use of a purely stoichiometric model of the host's metabolic network, OptStrain then predicts the effect of novel enzymes in the pathway, as well as which host genes should be up- or down-regulated in order to increase the production yield.

Criteria for ranking pathways

In *de novo* pathway engineering, it may sometimes be desirable to search for pathways before choosing a suitable host. Therefore, the best pathways have to be chosen based on a few theoretical criteria. Not everyone agrees on the criteria to be used for this process of prioritization (**Figure 1**). As mentioned above, Cho et al. (2010) use five criteria, including organism specificity and pathway distance. According to extreme pathway analysis (Papin, Price, Palsson 2002), however, the length of the pathway does not influence production rate. Even so, the energetic costs of producing more enzymes should be taken into account when considering longer pathways, which still makes pathway distance a relevant parameter. Organism specificity can only be an applicable criterion if a host has already been chosen. It can be disputed whether it is more desirable that the pathway consists of enzymes specific for one particular organism; combining enzymes from different organisms selected for experimentally tested activities is actually likely to yield more effective compound production, as the most catalytically efficient combination of enzymes that is available can then be identified from these. Other criteria can be the theoretically achievable yield and the achievable activity, which are both difficult to predict without analysis of the metabolic network of the host.

One of the few factors that can be determined independently is the theoretical thermodynamic favourability. In most methods, the thermodynamic favourability is measured by a group contribution method, which measures the Gibbs free energy of formation of groups of atoms of the products and intermediates (Jankowski et al. 2008). These groups are added up to a total Gibbs energy of every reaction in the pathway. When the Gibbs energy of formation adds up to a negative value, the reaction is defined as thermodynamically favourable. Cho et al. (2010) went beyond only using Gibbs free energy of formation, also taking into account the fluctuation of Gibbs energy between the reactions in the pathway: the less fluctuation, the more thermodynamically favourable the pathway, since a product of each reaction in the pathway is a reactant for the next. Less fluctuation of Gibbs free energy along the pathway also reduces the accumulation of intermediates, thus avoiding potential adverse effects on the host organism.

All in all, the algorithms and tools described provide a very useful toolbox for the synthetic biologist. By smartly combining the advantages of several tools, the predictions have the potential to make possible the exploration of pathways, which are chemically more versatile and hopefully at least as effective as those found in nature.

Metabolic modelling in candidate host organisms

One cannot predict, construct and investigate a new metabolic pathway without taking the host organism into account, as every new pathway has to take its place within the overall topology of a large native metabolic network (Breitling, Vitkup, Barrett 2008). Competition with native pathways and metabolites, unpredicted side products and feedback loops are only some of the possible effects that the context of the host organism can have on the new pathway.

One approach to find a suitable host organism for a pathway is to look for an organism that already has most of the enzymes from the designed pathway present in its native metabolic network. In this way, fewer enzymes would have to be introduced in the organism and thereby the metabolic network would be disturbed less. This idea has been incorporated in the ranking system of Cho et al. (2010), which prefers pathways that consist of many enzymes originating from the same organism. However, there are also reasons why using a host with many usable native enzymes may not be the best option. For example, it would then also be more likely that in the given organism more pathways exist which compete with the pathway that is to be introduced. Partially knocking out such native pathways can be a solution to this problem if they are not essential to the host. Algorithms such as OptKnock/OptFlux (Burgard, Pharkya, Maranas 2003; Rocha et al. 2010) are available to supply the researcher with suggestions on which pathway to knock out, based on metabolic flux simulations. More ambitiously, orthogonal synthetic systems, e.g., for transcription or translation, can be used to insulate the synthetic pathways even more. Unnatural DNA and amino acids are now available to keep the production machinery of the desired compound separate from the host at many levels (An and Chin 2009; Dixon et al. 2010; Neumann et al. 2010; Neumann, Slusarczyk, Chin 2010; Wang et al. 2007), although the fundamental metabolic link will still need to be maintained.

Genome-scale metabolic modelling

One of the most important computational developments that is likely to facilitate true *de novo* pathway engineering in the future is the development of genome-scale metabolic models (Durot, Bourguignon, Schachter 2009; Feist et al. 2009; Oberhardt, Palsson, Papin 2009; Price, Reed, Palsson 2004). Such models allow *in silico* prediction of the behaviour of a pathway within a candidate host organism using constraint-based flux balance analysis (Edwards, Ramakrishna, Schilling 1999; Orth, Thiele, Palsson 2010). In this approach, a steady-state flux distribution of the metabolic network is predicted based on the stoichiometry of each reaction, mass-balance constraints and an objective function that specifies towards which goal the fluxes are optimized. In traditional studies, the objective function often is the maximization of biomass production, while in the analysis of genetically engineered microbes a common alternative objective is the “minimization of metabolic adjustments (MOMA)”, finding the feasible flux distribution in the engineered strain that is closest to

the wild type situation (Segre, Vitkup, Church 2002). Flux balance analysis is often used effectively to increase product titers. For example, vanillin production in baker's yeast was increased twofold by selecting genes for knockout constructs based on flux balance analysis simulations (Brochado et al. 2010). In a similar fashion, Asadollahi et al. (2009) achieved 85% titer improvement after constructing model-guided knockouts. The more recently developed Thermodynamic Metabolic Flux Analysis (TMFA; Henry, Broadbelt, Hatzimanikatis 2007) adds further thermodynamic constraints based on the Gibbs free energy change of each reaction and the concentrations of metabolites, to identify only thermodynamically feasible flux distributions and thereby increase the predictive power.

Besides the MATLAB-based COBRA toolbox (Becker et al. 2007), which has become a near-standard in the field, TMFA can also be performed with user-friendly web servers or graphical user interfaces (GUIs), such as SurreyFBA (Gevorgyan et al. 2011), CycSim (Le Fevre et al. 2009) and the BioMet Toolbox (Cvijovic et al. 2010). The advantage of CycSim and BioMet is that they are web-based and therefore require no installation, but they are limited to the analysis of a small number of model organisms. The stand-alone tool SurreyFBA may have a somewhat steeper learning curve, but it is more complete and can be accessed both through a GUI and through (scripting from) the command-line. One of the major bottlenecks in the use of metabolic models is the low level of standardization of the SBML files that are used to store genome-scale models, sometimes making models incompatible between tools.

As essential as it might be, computational prediction of the effects of introducing a pathway into a host is a procedure that is still in the pioneering stage. In one early example, the previously discussed pathway prediction system BNICE was used to find novel pathways for biodegradation of the pollutant 1,2,4–trichlorobenzene by *Pseudomonas putida*, a known pollutant degrader (Finley, Broadbelt, Hatzimanikatis 2010). To predict which pathway was most effective and what would be the growth rate of the host with the implemented pathway, TMFA was applied to a genome-scale model of the metabolic network of the host. By implementing thermodynamic constraints in the modelling, the number of candidate pathways was reduced around 200-fold. This approach seems to be most useful when testing not-yet-existing pathways generated by pathway prediction algorithms.

Around forty manually curated genome-scale metabolic reconstructions of different bacteria have been published until now (Feist et al. 2009). Interestingly, a new approach has recently been published which can automatically generate metabolic models based on genome sequences, by annotating them in a uniform fashion, linking predicted enzymes to reactions, and filling in gaps (Henry et al. 2010). Although the models require subsequent manual curation, this methodology now provides the intriguing option of reconstructing multiple models in parallel and performing comparative analyses between them (Alam et al. 2011). The fact that the reconstruction method is uniform removes the large annotator bias which otherwise makes manually constructed models difficult to compare. Such models can be utilized to predict the suitability of the metabolic network topologies of multiple candidate hosts for production of a specified compound *in silico*, by introducing the pathway into these models, performing TMFA with a dual objective consisting of the biomass and the production of the compound, and comparing the predicted maximized fluxes to the target compound.

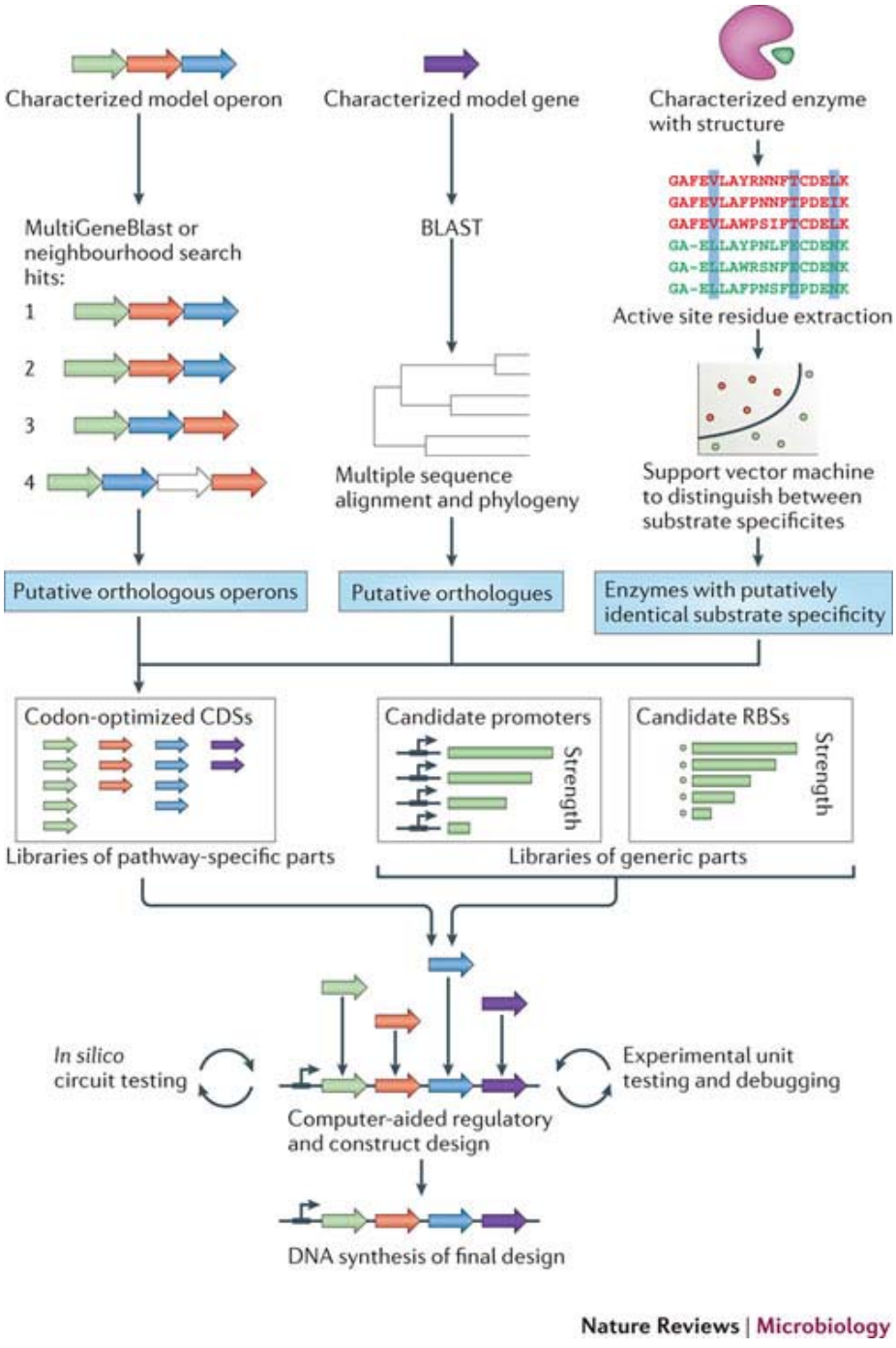

Figure 2. Scheme showing the steps involved in the identification of various parts, their refactoring, and their integration into transcriptional units. At the top, three strategies are shown for the identification of libraries of variants of pathway-specific genetic parts based on genetic and biochemical knowledge. The left panel shows how operons or gene clusters that are homologous to a characterized operon or gene cluster can be detected using neighbourhood orthology analysis (as in IMG (Mavromatis et al. 2009) or MultiGeneBlast (Medema, Takano, Breitling 2013)). The middle panel shows the well-accepted procedure for the identification of orthologues of a characterized gene, by homology search, multiple sequence alignment and phylogenetic tree construction. The right panel shows the method recently pioneered by Röttig et al. (2010) to identify enzymes with identical substrate specificity to a model enzyme if this enzyme is part of an enzyme family that contains multiple specificities. The method proceeds by automatic extraction of active site residues based on a crystal structure, training of a support-vector machine to distinguish between the different substrate specificity variants within the enzyme family, and classification of all homologs to identify those enzymes that have the desired specificity. Coding sequences (CDSs) of the identified pathway-specific parts are codon-optimized using e.g. OPTIMIZER (Puigbo et al. 2007), GeneDesign (Richardson et al. 2010) or Gene Designer (Villalobos et al. 2006). Libraries of generic parts are then acquired.

Host-specific promoters are collected, and RBSs are obtained with the aid of, e.g., the RBS Calculator (Salis, Mirsky, Voigt 2009; Salis 2011) and/or RBSDesigner (Na and Lee 2010). Both generic and pathway-specific parts are then used in computer-aided design of genetic constructs encoding the target biochemical pathway. After extensive *in silico* and *in vivo* testing and debugging, oligos are designed (e.g. using TmPrime (Bode et al. 2009) or DNAWorks (Hoover and Lubkowski 2002)) to synthesize the final design.

The outputs of *in silico* TMFA experiments are often complex and difficult to interpret in the context of the whole metabolic map of an organism. Fortunately, powerful visualization tools have recently been made available which can be used to colour pathways according to predicted fluxes. iPath2 (Yamada et al. 2011), for example, can generate coloured KEGG-based metabolic maps based on simple tables of EC numbers and corresponding colours. Interestingly, GLAMM (Bates, Chivian, Arkin 2011), a similar web service, offers the additional possibility to automatically highlight routes between two compounds in the metabolic map. Such visualization tools are indispensable for gaining a rapid understanding of the large output tables that are usually obtained through modelling. Combined with the user-friendly TMFA simulation tools mentioned above, they make this complex technique accessible to the non-specialist experimental microbiologist.

Strategies for parts identification

In order to construct a pathway, one will of course need to find the parts that can carry out the proposed enzymatic and regulatory steps, in the form of gene domains, genes, and operons (**Figure 2**). A range of parts is already available in online parts registries (Galdzicki et al. 2011), and several major efforts are under way to systematically characterize and standardize large numbers of biological parts in a consistent manner (Canton, Labno, Endy 2008). However, these parts registries are currently much more suitable for finding regulatory elements than for finding coding sequences of biosynthetic enzymes, as these parts are much more unique and specified. Therefore, effective genome mining of enzymatic parts is a crucial step in the construction and optimization of these parts.

To optimize metabolic fluxes through the candidate pathways, one needs to find the optimal combination of enzymatic parts in terms of individual catalytic efficiency as well as overall pathway reaction stoichiometry. Importantly, these parts do not necessarily have to originate from one existing natural system in one particular organism; instead, one could select a range of candidate parts from many different sources, characterize them (if possible, using some high-throughput assay) and integrate the codon-optimized versions of the most optimal combination of parts into a new pathway (see **Box 1**). To find combinations of enzymes that cooperate effectively, full exploitation of the genomic databases to make an optimal selection of all available parts is probably necessary.

The optimal computational strategies for harvesting potential parts from the genomic databases differ according to the nature of the parts, whether they are domains, genes or operons.

Identification of genes and domains

Gene domains can usually most easily be identified by traditional searching for the encoded protein

domains with curated profile Hidden Markov models from the Pfam (Finn et al. 2010), SMART (Letunic, Doerks, Bork 2009) or TIGRFAM databases. This can be done manually using the HMMer package (Eddy 2009) or using the SMART/PFAM architecture searches available on the web. After detection of the total set of domains, a high-priority subset can be identified by generating a phylogenetic tree and identifying the branches which specifically contain curated entries of domains with the desired enzymatic activity. If certain active site residues are known to be essential for this activity, their presence can be verified for the entries in multiple-sequence alignments of the selected branches. Alternatively, active sites can also first be predicted *ab initio* based on sequence conservation and/or structural information (Bray et al. 2009; Goyal, Mohanty, Mande 2007). If the desired parts are individual genes, candidates can be identified by a simple BLAST search, and analysed according to their phylogeny and active site residues in the same fashion as described above. As an additional criterion, the genomic context of candidate parts may be studied to verify the likelihood of the gene having the desired function in the context of the whole operon in which it is located. Finally, databases such as KEGG can be used to also find isoenzymes, which might catalyse the same reaction, even if they have no significant sequence homology. To increase the number of potential parts that can be tested even further, the set of identified genes can be supplemented by predicted evolutionary intermediates or ancestor genes or domains (Hall 2006; Thornton 2004) of different taxonomic branches which have a high potential to represent the desired parts.

In some cases, a manual inspection of active site residues of candidate gene or domain parts may not be sufficient to reliably predict their enzymatic activity. This is especially the case if the desired enzyme is a member of a broader enzyme family that encompasses multiple substrate specificities. More elaborate *in silico* approaches to find enzymes with the right substrate specificity or to identify the right mutations to get those enzymes will then be needed. Recently, an automated method was developed for classifying active sites from enzymes throughout an enzyme family using support-vector machines trained on sets of residues around the active site of enzymes with known substrate specificities from within the family (Röttig, Rausch, Kohlbacher 2010). The approach has already been successfully implemented to predict the extremely variable substrate specificities of non-ribosomal peptide synthetases (RW.ERROR - Unable to find reference:583).

Identification of multigene modules

Often, the desired parts are not single domains or genes but complete operons or gene clusters, e.g. the biosynthetic operon for a precursor needed to produce the compound of choice. In such cases, one will want to avoid having to combine the results of a large number of individual BLAST searches manually to find the homologous genomic regions. One currently available tool for identifying genomic regions homologous to a certain model operon is the 'identify homologous regions' tool in the Integrated Microbial Genomes (IMG) system of the Joint Genome Institute (Mavromatis et al. 2009). However, an important disadvantage of this tool is that it does not cover all information stored in genomic databases such as GenBank. Additionally, it is difficult to specifically look for overall homologs of genetic elements at the operon or gene cluster levels, as the search is always performed on the neighbourhoods of a single gene. If the biosynthetic operon is always part of a specific type of secondary metabolite biosynthetic gene cluster, the comparative gene cluster analysis module from antiSMASH (Medema et al. 2011b) can be used to find gene clusters that have the same operon. Additionally, to be able to very specifically look for all genomic regions homologous to a given query operon or gene cluster, a new tool, MultiGeneBlast, has been

developed in our group (Medema, Takano, Breitling 2013). This tool effectively combines the individual BLAST results of the various genes in the query gene cluster to rank all genomic regions from GenBank based on the number of BLAST hits from the query gene cluster, the conservation of gene order and the cumulative BLAST bit score.

Refactoring the parts and designing transcriptional units

When raw candidate parts have been screened and an optimal combination has been selected, one will still need to optimize the sequences for the targeted host organisms, and combine them into transcriptional units with a well-designed regulatory circuitry (**Figure 2**). Several algorithms have been developed to aid in these processes. Additionally, the development of drag-and-drop computer-aided design (CAD) approaches (Marchisio and Stelling 2009) makes the life of the biological designer much easier.

Computer-aided design software

When trying to heterologously express a genetic construct that consists of non-native parts, a crucial factor in obtaining high protein expression rates is optimizing the codon usage in the coding regions, by matching it to the more abundant tRNA species in the host. A range of tools has been developed which can use a codon frequency table of the target host organism to optimize the codon usage of a given protein-coding sequence (CDS). Arguably most straightforward in its use is Optimizer (Puigbo et al. 2007), a simple web tool which does exactly this. Another web server, GeneDesign (Richardson et al. 2010), offers several further options such as the addition of restriction sites. For advanced users, there is Gene Designer 2.0 (Villalobos et al. 2006), which offers a comprehensive drag-and-drop user interface for the construction of genetic constructs that also include regulatory parts such as promoters, ribosome binding sites (RBSs) and transcriptional terminators. Another CAD-design tool is GenoCAD (Czar, Cai, Peccoud 2009), which is less comprehensive but generally easier to handle and provides an additional interesting feature: it assists the user in correctly designing genetic constructs by interactively checking it against a user-specified set of grammatical rules (e.g., a transcription unit is always composed according to the following pattern: promoter— [RBS—CDS]$_n$—terminator). To design a simple *E. coli* three-gene operon construct with GenoCAD, for example, one would first select the *E. coli* grammar with the parts libraries that contain parts matching to this grammar. Then one would arrange the desired architecture in a way that complies to the grammar (promoter-RBS-CDS-RBS-CDS-RBS-CDS-terminator), and finally select the specific parts for each entry from the libraries.

Regulatory circuits are crucial for designing well-balanced genetically engineered machines, and much synthetic biology efforts have focused on developing these (Mukherji and van Oudenaarden 2009; Tamsir, Tabor, Voigt 2011). In order to design regulatory circuits and test their outputs, several CAD tools are available that have integrated simulation capacities to predict whether the envisioned regulatory mechanisms will function as intended. SynBioSS (Weeding, Houle, Kaznessis 2010), for example, simulates regulatory circuits by creating network models of reactions that represent transcription, translation, *cis* and *trans* regulatory effects, and degradation. Of specific interest for

pathway design is BioNetCad (Rialle et al. 2010), which allows a similar kind of simulation of the logic inherent in biochemical networks consisting of enzymes and metabolites.

Designing regulatory parts

Obviously, optimized regulatory parts that are to be used in genetic circuits and constructs are likely not to be available and will therefore usually have to be designed for the intended usage. Synthesizing and characterizing host-specific libraries of characterized RBSs and sigma-factor binding sites (SFBSs) will be useful in many cases, as regulatory parts with the correct binding strength can then be selected from these for incorporation into the genetic construct. Yet even if this is done, the strength of an RBS (and probably of an SFBS as well) partially depends on its DNA sequence context (Salis, Mirsky, Voigt 2009). Therefore, some RBSs may not function as expected in their new context. To find a suitable RBS in such cases — or in cases when no RBS library is available — the online tools RBS calculator (Salis, Mirsky, Voigt 2009; Salis 2011) or RBS designer (Na and Lee 2010) can be used to suggest an RBS sequence for a given desired translation initiation rate. These algorithms are based on thermodynamic models of the molecular interactions between the ribosome complex and mRNA transcripts, and use these to predict a translation initiation rate. Although the RBS calculator does not necessarily succeed in reaching a global optimum, it enables rapid design of RBS sequences that are sufficient for most purposes. It would be very helpful if a similar tool would be generated for SFBS design in the near future.

DNA synthesis and integration

When an adequate design has been obtained, the next step will be the synthesis of the DNA parts. Online algorithms — notably DNAWorks (Hoover and Lubkowski 2002) and TmPrime (Bode et al. 2009) — are available to design oligonucleotides for PCR-based gene synthesis, which also include integrated codon optimization functionalities. These can be highly important for efficient heterologous expression of a gene. For example, in terpenoid-producing *E. coli* strains, Chang et al. (2007) reported a 2.5-fold increase in production titers after codon optimization. Several recent developments in DNA synthesis (Kosuri et al. 2010; Matzas et al. 2010; Quan et al. 2011) now allow DNA synthesis in a throughput that is sufficient even to construct thousands of variants of any DNA parts, which can then be tested to pinpoint the ones which function best. These developments pose new opportunities for computational tools, for example to allow design of libraries of parts variants that optimally cover the relevant parts of sequence space.

In order to arrive at a functional design for the DNA construct encoding a biosynthetic pathway, significant lessons can be learned from programming. For example, 'unit testing' — the functional testing and debugging of every individual component before putting everything together— is crucial to keep the complexity of the debugging process manageable. Fluorescent or other biological signals may be introduced into the construct to mimic the 'print statement' that is often used in computer software debugging to test the successful execution of a programmatic unit. When one has arrived at a functional design, it can be inserted into the chromosome of a specified plug-and-play host (Medema et al. 2011a) or in a multigene expression plasmid (Heneghan et al. 2010).

Future perspectives

In general, the ability to computationally predict pathways, identify variants of the necessary parts, and model them in genome-scale metabolic networks of candidate host organisms offers great promise to accelerate the developing field of synthetic pathway engineering. In this area, systems biology can inform and complement synthetic biology approaches, which could lead to important breakthroughs in the near future (**Figure 3**).

The ambition of synthetic biology is to design biological systems based on first principles, irrespective of which combinations of parts happen to be used in nature. The algorithms that have been developed in recent years are likely to facilitate reaching this goal. However, due to the immense challenge of biological complexity (Kwok 2010), the development of additional algorithms specifically focused on the needs of synthetic biology projects will be crucial to allow the field to mature.

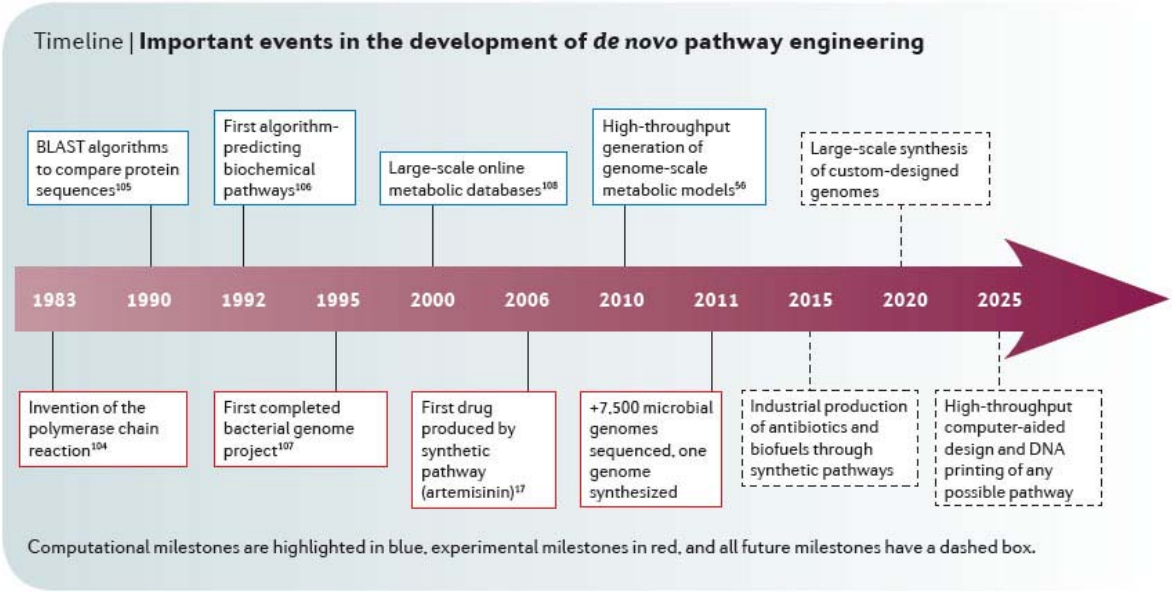

Figure 3. Time line of important developments related to *de novo* pathway engineering, highlighting experimental (yellow) and computational (green) milestones. 1983: Invention of the polymerase chain reaction (Mullis 1990). 1990: Development if the BLAST algorithm (Altschul et al. 1990). 1992: Construction of the first algorithm to generate biochemical pathways from a database (Mavrovouniotis, Stephanopoulos, Stephanopoulos 1992). 1995: First bacterial genome sequence, of *Haemophilus influenza* (Fleischmann et al. 1995). 2000: Rise of online metabolic databases (such as KEGG; Kanehisa and Goto 2000). 2006: Keasling and co-workers (Ro et al. 2006) report the production of artemisinic acid in engineered yeast. 2010: First pipeline for high-throughput generation of metabolic models (Henry et al. 2010). 2011: Sixteen years after the first bacterial genome sequence, approximately 7500 prokaryotic genome sequences are available. In red, possible future milestones of synthetic microbiology are forecast.

In all key aspects of synthetic microbiology, the speed of progress will largely depend on the headway made in software development and data systematization. Fortunately, there are great opportunities for computational breakthroughs all around.

In the field of metabolic modelling, the advances in high-throughput generation of genome-scale models (Henry et al. 2010) is likely to inaugurate a whole new era. A field of comparative modelling (Alam et al. 2011) will develop, in which the metabolic network topologies from a range of genome-scale models can be compared rapidly. Once algorithms are developed to automatically add one or a

few enzymes to each model to simulate growth on a certain growth medium, the approach will be a useful way to compare for a range of medium types, which organisms have network topologies that are suitable for growth and productivity. Also, when a specific host organism has already been chosen, the same algorithm could be implemented to aid in medium design. Ideally, the design of synthetic pathways and growth media would be an integrated process in which the pathways are supplemented with the enzymes necessary to arrive at an optimal combination of medium and pathway. The field of metabolomics, which has its own crucial branch of software development (Wishart 2009), will be key in coupling the predictions to actual experimental measurements.

The available pathway prediction tools will need to be made more user-friendly (with GUIs or web servers) and linked to well-curated databases of experimentally characterized enzymatic parts in an integrated framework. During the last decades, an enormous range of enzymes have been characterized that are involved in the synthesis or tailoring of small molecule scaffolds (Walsh and Fischbach 2010), but little systematic data archiving has been performed thus far. For such purposes, databases like KEGG provide sparse information, being largely focused on primary metabolism. Besides enzymes, it would be very helpful if transporters for small molecules and resistance genes against toxins or antibiotics would also be categorized systematically in a database, and linked to the chemical structure of the corresponding small molecule. Algorithms for chemical (sub)structure similarity searching (Ridley 2001; Willett, Barnard, Downs 1998) could then be employed to rapidly search for enzymes or transporters that are likely to synthesize or transport a compound of choice, or which would otherwise have a functionality that is related closely enough to be modified successfully using, e.g., directed evolution.

It is also critical that computational developments go hand-in-hand with developing experimental approaches. High-throughput transcription factor binding affinity characterization using protein binding microarrays (Berger et al. 2006) or microfluidics (Fordyce et al. 2010) could for example be perfectly linked up to algorithms for the design of transcription factor and sigma factor binding sites (van Hijum, Medema, Kuipers 2009). Alternatively, when a close homologue of a desired transcription factor has already been characterized in a related species, phylogenetic footprinting approaches (Francke et al. 2008) can be employed to automatically predict species-specific binding motifs of the orthologue which one wants to utilize. Yet when a synthetic transcription factor is heterologously expressed, algorithms would probably still be necessary to correct for the difference in GC content between the species, which is likely to influence the heterologous binding dynamics on a chromosome which is foreign to those for which experimental data have been obtained. Finally, with the expected advances in synthetic genomics, the development of algorithms for optimal integration of genetic constructs into a chromosomal design will be highly important to arrange both operon-level genetic organization (Lim, Lee, Hussein 2011) and higher-level organization (Montero Llopis et al. 2010) for optimal gene expression levels.

When computational and experimental breakthroughs thus go hand-in-hand in an integrated manner, synthetic pathway engineering may well become one of the major drivers of applied synthetic biology.

Part II

Computational genomic analysis of microbial secondary metabolism

Chapter 8

The sequence of a 1.8-Mb bacterial linear plasmid reveals a rich evolutionary reservoir of secondary metabolic pathways

Published as:

M.H. Medema, A. Trefzer, A. Kovalchuk, M. van den Berg, U. Müller, W. Heijne, L. Wu, M.T. Alam, C.M. Ronning, W.C. Nierman, R.A. Bovenberg, R. Breitling, E. Takano (2010) The sequence of a 1.8-Mb bacterial linear plasmid reveals a rich evolutionary reservoir of secondary metabolic pathways. *Genome Biology and Evolution* 12: 212-224.

Abstract

Plasmids are mobile genetic elements that play a key role in the evolution of bacteria by mediating genome plasticity and lateral transfer of useful genetic information. While originally considered to be exclusively circular, linear plasmids have also been identified in certain bacterial phyla, notably the actinomycetes. In some cases, linear plasmids engage with chromosomes in an intricate evolutionary interplay, facilitating the emergence of new genome configurations by transfer and recombination or plasmid integration.

Genome sequencing of *Streptomyces clavuligerus* ATCC 27064, a Gram-positive soil bacterium known for its production of a diverse array of biotechnologically important secondary metabolites, revealed a giant linear plasmid of 1.8 Mb in length. This megaplasmid (pSCL4) is one of the largest plasmids ever identified and the largest linear plasmid to be sequenced. It contains more than 20% of the putative protein-coding genes of the species, but none of these is predicted to be essential for primary metabolism. Instead, the plasmid is densely packed with an exceptionally large number of gene clusters for the potential production of secondary metabolites, including a large number of putative antibiotics, such as staurosporine, moenomycin, β-lactams and enediynes. Interestingly, cross-regulation occurs between chromosomal and plasmid-encoded genes.

Several factors suggest that the megaplasmid came into existence through recombination of a smaller plasmid with the arms of the main chromosome. Phylogenetic analysis indicates that heavy traffic of genetic information between *Streptomyces* plasmids and chromosomes may facilitate the rapid evolution of secondary metabolite repertoires in these bacteria.

Introduction

Prokaryotic chromosomes were originally thought to lack many of the features that characterize eukaryotic chromosomes, such as linearity and the possession of telomeres (Bendich and Drlica 2000). However, during recent decades a few bacterial taxa have been shown to have linear chromosomes which contain telomere-like structures (Hinnebusch and Tilly 1993). Among these are the high-GC, Gram-positive, soil-dwelling *Streptomyces* bacteria (Chen et al. 1994; Hopwood 2006), renowned for their capacity to produce a vast array of natural products. Their linear chromosomes are relatively long (8–10 Mb) and consist of a conserved 'core' of 5–6 Mb and variable 'arm' regions (Hopwood 2006). Nearly all genes that are likely to be unconditionally essential are located within the 'core.' Interestingly, besides possessing linear chromosomes, streptomycetes also often contain linear plasmids, extrachromosomal DNA molecules that replicate independently from the main chromosome and contain their own telomere-like structures (Chater and Kinashi 2007; Chen 2007). Evidence is accumulating that the evolution of streptomycete chromosomes and plasmids can be very dynamic, because of their relative instability (Chen et al. 2002; Volff and Altenbuchner 1998; Widenbrant et al. 2007).

Substantial effort in the study of streptomycetes is focused on the secondary metabolites they produce. Many bacterial secondary metabolites or their derivatives are used as antimicrobial agents, while some are used as antitumor drugs, immunosuppressive agents or cholesterol-lowering drugs (Bode and Müller 2005; Gullo et al. 2006; Newman and Cragg 2007). More than half of all known

antibiotics originate from the streptomycetes (Berdy 1995; Challis and Hopwood 2003), and many more clearly remain to be discovered. A statistical analysis of secondary metabolism has predicted that streptomycetes may have the capacity to produce as many as 10^5 secondary metabolites (Watve et al. 2001). The finding that a single genus can carry such a massive number of secondary metabolite biosynthetic gene clusters is correlated with the location of many of them in the arm regions of the chromosome, which are extremely variable between species. In a few cases, such clusters have been shown to be plasmid-borne. Thus the 365-kb plasmid SCP1 of *Streptomyces coelicolor* carries the biosynthetic gene cluster for methylenomycin A (Bentley et al. 2004; Kirby and Hopwood 1977), and the 210-kb plasmid pSLA-2 of *Streptomyces rochei* carries four secondary metabolite biosynthetic gene clusters: three polyketide synthase (PKS) clusters (for lankacidin, lankamycin and a mithramycin-like compound), and a carotenoid biosynthetic cluster (Mochizuki et al. 2003).

One of the most important industrial streptomycete species is *Streptomyces clavuligerus*. Multiple natural products from this organism have been structurally and functionally characterized, and two of these are used in the clinic, the β-lactam antibiotic cephamycin C (Alexander and Jensen 1998) and the β-lactamase inhibitor clavulanic acid (Saudagar, Survase, Singhal 2008). Clavulanic acid is widely used in combination with the semi-synthetic β-lactam amoxicillin to treat various bacterial infections (Brogden et al. 1981). The biosynthetic genes for both compounds are encoded by a single intensively studied gene supercluster (Liras, Gomez-Escribano, Santamarta 2008). Several clavam antibiotics closely related to these β-lactam ring-containing compounds are also produced by *S. clavuligerus*, and are encoded by separate gene clusters (Evans et al. 1983; Tahlan, Anders, Jensen 2004; Tahlan, Park, Jensen 2004; Tahlan et al. 2004; Tahlan et al. 2007; Zelyas et al. 2008). At least three additional antibiotics have been reported to be produced by *S. clavuligerus*, although no biosynthetic genes for these have been described: the pyrrothine-class antibiotic holomycin (Kenig and Reading 1979; Oliva et al. 2001); a tacrolimus-like macrolide (Kim and Park 2008); and a compound related to the nucleoside antibiotic tunicamycin (Kenig and Reading 1979).

Here, we describe the genome sequence of *S. clavuligerus* ATCC 27064 and show that this species has a unique 1.8-Mb linear megaplasmid, which is densely packed with 25 putative secondary metabolite gene clusters, in addition to its 6.8-Mb chromosome, which also contains 23 such clusters. Some of these clusters strongly resemble known gene clusters, others appear to be completely novel, and a number show extravagant features that have never been observed before. The megaplasmid found in *S. clavuligerus* is by far the largest linear megaplasmid ever sequenced, and its gene complement sheds light on the rapid and dynamic evolution of secondary metabolite repertoires in bacteria, in addition to being a rich and compact potential source of novel bioactive metabolites.

Materials and methods

Genome sequencing and assembly

The genome of *S. clavuligerus* ATCC 27064 was sequenced and assembled by random shotgun sequencing. Sanger sequencing of shotgun libraries with insert sizes of 3 kb, 10 kb, and 50 kb was accomplished using ABI 3700 sequencers as described by (Venter et al. 2001). Sequences were

assembled using Celera Assembler (Levy et al. 2007). Genome assembly was facilitated and validated with an optical restriction map (Latreille et al. 2007) acquired from OpGen (Madison, Wisconsin). Finally, as many gaps as possible were filled using sequences from a cosmid library created in-house at the DSM Biotechnology Center. We estimate the final genome assembly to be > 99.7 % complete.

Gene prediction, functional annotation and comparative analysis

Straightforward gene prediction by Glimmer3 (Delcher et al. 2007) appeared to give a large number of false positives and negatives. The high GC content of *Streptomyces* species leads to the presence of many long open reading frames (orfs) which are not actual genes but only exist by coincidence because of the low number of AT-rich stop codons. These long reading frames often are not actual genes. Therefore, we complemented Markov model-guided prediction by Glimmer3 with mapping of BlastP searches of all putative orfs to the non-redundant protein database (using an in-house Python script), and with analysis of the GC frame plots in Artemis (Rutherford et al. 2000) to manually identify all putative genes accurately.

tRNA and tmRNA genes were identified using ARAGORN, and rRNAs were identified using the RNAmmer tool. For functional annotation, all proteins were first processed with the AutoFACT automatic annotation pipeline (Koski et al. 2005) through hierarchical information transfer from the best hits in the Uniref90, NCBI nr, COG, KEGG, CDD, PFAM, and SMART databases. A round of manual re-annotation followed to ascertain the accuracy of the annotation.

Genome alignment of the DSM and Broad Institute draft genomes was performed using the Mauve genome alignment program (Darling et al. 2004). Clusters of orthologous groups were constructed using the OrthoMCL software (Li, Stoeckert, Roos 2003), using cut-offs of e-value < 1E-05 and identity > 60%. Dotplots for genome comparison were created with Gepard (Krumsiek, Arnold, Rattei 2007). GC skew plots were created using GenSkew (http://mips.gsf.de/services/analysis/genskew/).

Phylogenetic analysis of plasmid-encoded transposase proteins was performed as follows. The *S. clavuligerus* transposase sequences were blasted against the non-redundant (nr) protein database with BlastP. For each query sequence, the 25 best blast hits were aligned using Muscle 3.6 (Edgar 2004), the alignment was quality-trimmed with Gblocks 0.91b (Talavera and Castresana 2007), and an approximate maximum likelihood phylogenetic tree was generated using FastTree 2.1.1 (Price, Dehal, Arkin 2010). Finally, all the hits incorporated in the trees were blasted against a self-assembled database of *Streptomyces* chromosomes to identify transposases from *Streptomyces* chromosomal termini (distance <1 Mb to either end of the chromosome) using BlastX. Query sequences present in a monophyletic taxon comprised of at the most three proteins, in which at least one of the two other proteins was a transposase from a *Streptomyces* chromosomal terminus, were considered as positive hits.

Metabolic model construction

Published genome-scale metabolic models of *S. coelicolor*, *Escherichia coli* and *Saccharomyces cerevisiae* (Borodina, Krabben, Nielsen 2005; Duarte, Herrgård, Palsson 2004; Feist et al. 2007) were

used together with a genome-scale model of *Penicillium chrysogenum* (DSM, unpublished data) as a reference to construct the metabolic model of *S. clavuligerus*. Gene-reaction associations were putatively assigned to *S. clavuligerus* genes based on sequence homology to genes included in the reference genome-scale metabolic models. 'Dead-end' metabolites (i.e. metabolites that are neither consumed nor produced under any conditions) were identified, and gaps were closed, if possible, by BLAST searches on the *S. clavuligerus* genome with enzymes from the KEGG database that could connect dead-end metabolites. The gene-reaction associations were further manually verified and curated using the EC numbers from the functional annotation to correct for erroneous gene-reaction associations. The metabolic model was checked for biomass formation *in silico* under minimal growth medium conditions (glycerol, ammonia, phosphate and sulfate) using the COBRA toolbox (Becker et al. 2007). An SBML version of the final model is available from the authors upon request.

Accession of genome sequence information

The Whole Genome Shotgun project of *S. clavuligerus* strain ATCC 27064 has been deposited at DDBJ/EMBL/GenBank under the accession ADGD00000000. The version described in this paper is the first version, ADGD01000000.

Results and Discussion

Architecture of the Streptomyces clavuligerus genome

Pulse-field gel electrophoresis of restriction-digested fragments in the nineties lead to an estimated size of 6.8 Mb for the chromosome of *S. clavuligerus* (Chen et al. 1994). Our genome sequencing and assembly confirmed this estimate. However, remarkably, a second large contig of 1.8 Mb was identified as well.

The optical restriction map that we obtained indicated that this second contig represents a separate replicon. This was confirmed by almost perfectly symmetrical cumulative GC skew plots of both contigs (**Supplementary Figure 1**), characteristic for bacterial chromosomes and plasmids (Grigoriev 1998). Origins of replication (Oris) were identified near the peak of the plots for both replicons on the basis of homology.

The identification of the two large scaffolds was later independently confirmed by a pre-publication draft sequence from the Broad Institute (GenBank accession number ABJH00000000), although our assembly of the 6.8 Mb replicon differed from theirs by a large inversion at virtually identical rRNA operons (**Supplementary Figure 2**). The optical restriction map that we obtained and our GC skew analysis indicated that our assembly is most probably correct.

The discovery of a second large replicon in *S. clavuligerus* is not altogether surprising, as in 1978 Kirby suspected the presence of a plasmid when he found that a genetic element without linkage to the chromosome influenced holomycin production (Kirby 1978). Intriguingly, pulse-field gel electrophoresis in the closely related *S. clavuligerus* strain NRRL 3585 — the type strain which should be identical to strain ATCC 27064 — has shown the presence of three plasmids (pSCL1–3) ranging in

size between 11 kb and 430 kb (Netolitzky et al. 1995). However, only a 7-kb replicon matching to the smallest of these plasmids — the 11-kb plasmid pSCL1 (Wu and Roy 1993) — was identified in our assembly. No replicons of the sizes reported for plasmids pSCL2 or pSCL3 were identified in our assembly, and the nucleotide sequence of pSCL2 (Wu et al. 2006) did not match to any part of our contigs in a BlastN analysis. This implies that the evolution of streptomycete genomes by plasmid acquisition, loss, or recombination is extremely dynamic.

We identified 7281 putative protein-encoding genes on the two large replicons, as well as six ribosomal RNA (rRNA) operons, 72 transfer RNA (tRNA) genes and 14 pseudogenes (**Table I**). Strikingly, the sum of all genomic features of the two replicons matches very well to the typical make-up seen in other *Streptomyces* genomes.

The 1.8-Mb replicon has the hallmarks of a giant linear plasmid

We set out to characterize and compare the two replicons. First of all, the central region of the 6.8-Mb replicon had very large regions of conserved synteny to the chromosomes of the four completed *Streptomyces* genomes, while the smaller replicon had no regions with significant homology to these chromosomes (**Supplementary Figure 3**). Moreover, the large replicon had an origin of replication typical of *Streptomyces* chromosomes, including *dnaA* and *dnaN* (SCLAV_2911–2912). The 6.8-Mb replicon thus appears to represent a typical *Streptomyces* chromosome. Interestingly, it is significantly smaller than other published *Streptomyces* chromosomes, which range in size between 8.5 and 9.0 Mb (Bentley et al. 2002; Ikeda et al. 2003; Ohnishi et al. 2008) (**Table I**). Comparative analysis shows that the conserved core region of *Streptomyces* chromosomes is present in *S. clavuligerus*, yet the typical large chromosomal arms are not (**Supplementary Figure 3**).

As stated in the previous section, the smaller replicon appeared to have its own Ori, highly homologous to those of the *S. coelicolor* plasmid SCP1 and the *S. rochei* plasmid pSLA2-L, two *Streptomyces* plasmids that are known to harbor secondary metabolite gene clusters. The Ori lies at around 945 kb, adjacent to close homologues of the plasmid-type DNA primase / replication proteins Orf1 and Orf2 (SCLAV_p0889 and SCLAV_p0888) from the *S. coelicolor* SCP1 plasmid (Redenbach et al. 1999), and an additional DNA primase / polymerase (SCLAV_p0890).

Both replicons have genes that encode ParAB chromosome partitioning proteins near their predicted Ori (<10 kb). The *parAB* genes of the 6.8-Mb replicon (SCLAV_2901–2902) strongly resemble the *S. griseus* and *S. avermitilis* chromosomal *parAB* genes, while the *parAB* genes of the 1.8-Mb replicon (SCLAV_p0884–0885) are plasmid-type, and strongly resemble the *parAB* genes on the SCP1 plasmid of *S. coelicolor* A3(2) (**Supplementary Figure 4**).

The smaller replicon also has its own *tap* and *tpg* genes, which encode proteins necessary for telomere replication of *Streptomyces* chromosomes and plasmids (Bao and Cohen 2001; Bao and Cohen 2003). No *tap* or *tpg* genes were identified on the main chromosome, as for the *Rhodococcus sp.* RHA1 genome (McLeod et al. 2006). The chromosome therefore seems to depend on the *tap*/*tpg* genes of the smaller replicon. However, there could also be a different, unknown, system for telomere replication; the recent identification of the *tac*/*tpc* system (not present on the *S.*

clavuligerus chromosome) of the *S. coelicolor* SCP1 plasmid shows that the *tap*/*tpg* system is not universal for streptomycete chromosomes and plasmids (Huang et al. 2007).

Genes for conjugative transfer were also detected on both replicons. Again, while the chromosome carries typical chromosomal-type *traAB* genes (SCLAV_4235–4236), the smaller replicon has plasmid-type *traAB* genes (SCLAV_p0254–0255) similar to those of the *S. coelicolor* plasmid SCP2 (Brolle et al. 1993).

Based on these observations, we conclude that the 1.8-Mb scaffold has all the hallmarks of being a giant version of a typical *Streptomyces* linear plasmid.

The smaller replicon is a megaplasmid predicted to be dispensable for the core metabolism of S. clavuligerus

Extrachromosomal genetic elements of bacteria can have quite different functions compared with the chromosomes. In some cases, they encode only accessory functions, while in others they have evolved to encode essential cellular functions as well. Although megaplasmids and chromosomes are part of a continuous spectrum, Bentley and Parkhill have proposed to distinguish chromosomes from megaplasmids not by their replication genes, but by whether they are essential for growth of the organism and whether they carry ribosomal RNA operons (Bentley and Parkhill 2004). We therefore set out to predict whether the 1.8-Mb replicon is essential for growth of *S. clavuligerus*.

Strikingly, all stable RNAs necessary for primary metabolism (ribosomal and transfer RNAs) are encoded on the main chromosome (**Table I**). A few putative tRNAs are encoded on the 1.8-Mb replicon, but these appear functionally redundant compared to those on the chromosome according to their amino acid specificity and anticodon sequence. To investigate whether the protein-coding genes of the replicon are likely to be involved in essential cellular functions, a *Streptomyces* core genome (Ussery, Wassenaar, Borini 2008) of 976 genes was defined, with each gene representing a cluster of orthologous genes present in all four completed *Streptomyces* genomes. This core genome had only 14 significant blastP hits (E-value < 1E-05 and identity > 60%) to the 1.8-Mb replicon, distributed throughout its length, with none appearing to be required for crucial cellular processes (see **Supplementary Table I**). Furthermore, all 21 out of the 24 *Streptomyces*-specific signature genes (Ohnishi et al. 2008) that are present in the *S. clavuligerus* genome are located on the *S. clavuligerus* chromosome. Therefore, the 1.8-Mb replicon does not seem to encode any of the essential genetic complement of *S. clavuligerus* ATCC 27064. This is corroborated by the fact that almost all of the genes with the highest codon adaptation indexes (CAIs) and third codon GC percentages (GC3s) — which are likely to have housekeeping functions (Wu, Culley, Zhang 2005) — lie on the chromosome (**Supplementary Figure 5**).

In order to predict whether *S. clavuligerus* could grow normally using standard carbon and nitrogen sources without the help of enzymes encoded on the 1.8-Mb replicon, we constructed a metabolic model based on our functional annotation, comparisons with known metabolic models from other organisms (Borodina, Krabben, Nielsen 2005; Duarte, Herrgård, Palsson 2004; Feist et al. 2007), and subsequent gap-filling (**Table II**). This model successfully simulated growth of *S. clavuligerus* on various carbon and nitrogen sources. *In silico* knock-outs of all enzyme-coding genes specific to the

1.8-Mb replicon resulted in no growth defects when growth was simulated using glycerol as a carbon source. This suggests that no essential genes of primary metabolism lie on the megaplasmid.

Nonetheless, a few enzymes for primary metabolism are encoded on the 1.8-Mb replicon. Among them are malate synthase and isocitrate lyase, enzymes of the glyoxylate pathway required for bacteria to grow on acetate. However, there is a malate synthase isoenzyme gene on the chromosome, which has been experimentally verified before (Chan and Sim 1998), and copies of alternative actinomycete glyoxylate pathway genes (*ccr*/*meaA* and *mcl1* (Akopiants et al. 2006)) were also detected on the main chromosome, possibly making the plasmid-encoded copies redundant. Therefore, we predict that, even when growing on alternative carbon sources, the 1.8-Mb replicon is dispensable for primary metabolism. Experimental analysis by either curing of the whole replicon or constructing multiple targeted knock-outs should confirm this.

As the 1.8-Mb replicon does not seem to encode any functions essential to primary metabolism, the definition of Bentley and Parkhill supports the labelling of the 1.8-Mb replicon as a megaplasmid. As it is the fourth *S. clavuligerus* plasmid to be described, we name it pSCL4. It is the largest linear plasmid ever sequenced (Molbak et al. 2003). If the above definition is strictly applied, pSCL4 is actually the largest plasmid ever sequenced, as the only larger plasmid (the circular plasmid pGMI1000MP from *Ralstonia solanacearum*, (Salanoubat et al. 2002)) should then be classified as a chromosome: it contains an important part of the metabolic core (an rRNA locus with two tRNA genes, a gene coding for the alpha-subunit of DNA polymerase III, a gene for the protein elongation factor G, as well as several important enzymes of primary metabolism, including amino acid and cofactor biosynthesis), and thus is probably essential under many growth conditions (Salanoubat et al. 2002).

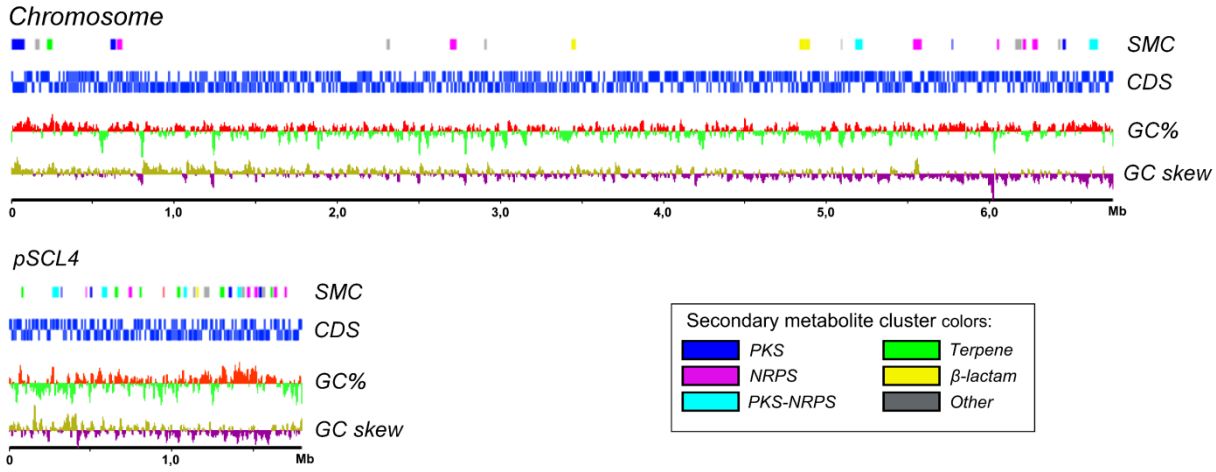

Figure 1: Overview of the putative secondary metabolism gene clusters in the *S. clavuligerus* genome. Chromosomal maps of the *S. clavuligerus* genome, showing both the chromosome and the megaplasmid pSCL4. SMC: secondary metabolite clusters, shown in different colors according to their type. CDS: coding sequences. GC%: GC percentage around the midline average of 72%. GC skew: GC skew, positive (brown) or negative (purple).

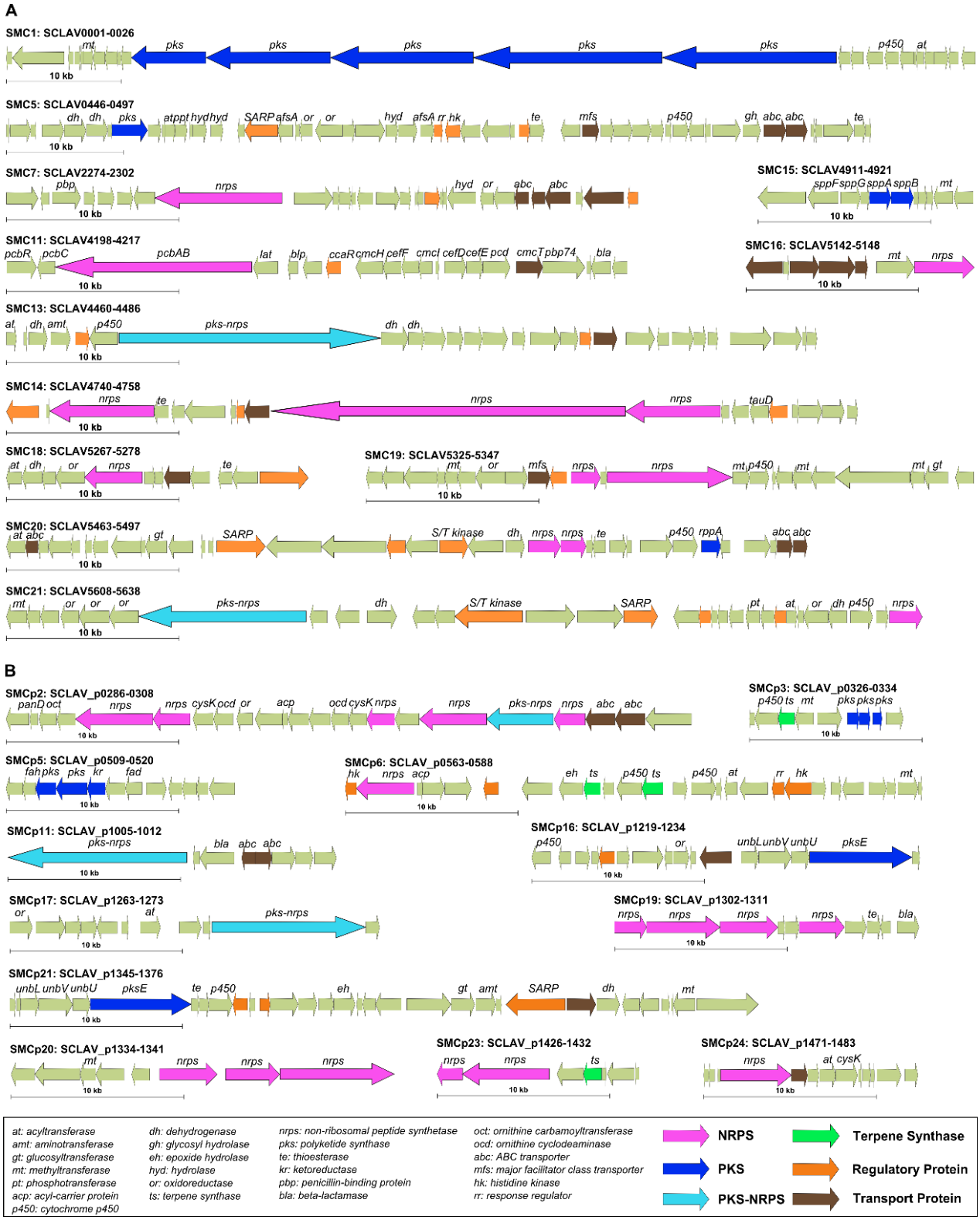

Figure 2: Secondary metabolite biosynthetic gene clusters: a detailed picture. Putative secondary metabolite gene clusters (SMCs) encoding non-ribosomal peptide synthetases and polyketide synthases on A) the chromosome and B) the megaplasmid pSCL4 of *S. clavuligerus* ATCC 27064. Different core and accessory genes are depicted in different colors as indicated. Cluster borders are approximate, and were estimated on the basis of functional annotations and putative operon structures. Details on the putative secondary metabolite gene clusters not shown in this figure (encoding beta-lactams, terpene synthases, lantibiotics, etc.) can be found in **Supplementary Table II**.

The megaplasmid is packed with secondary metabolite gene clusters

As the pSCL4 megaplasmid does not appear to harbor any housekeeping genes, we were interested in predicting the biological role of such a giant replicon. To our surprise, pSCL4 is packed with gene clusters putatively encoding the biosynthesis of secondary metabolites. No fewer than 25 such gene clusters, a number on the same order as observed in the chromosomes of other *Streptomyces* genomes, are dispersed throughout the plasmid (**Figure 1**). Together with the clusters identified on the chromosome, the total number of putative secondary metabolite gene clusters (SMC) identified in *S. clavuligerus* is 48, a number unprecedented in any bacterium (**Figure 2A&B**). These include ten putative NRPS gene clusters, eight putative PKS gene clusters and six gene clusters putatively encoding NRPSs as well as PKSs or NRPS-PKS hybrids, as well as 12 clusters putatively encoding one or more terpene synthases or cyclases. An overview of all clusters encoding putative NRPSs and PKSs is shown in **Figure 2**, and further details are given in **Supplementary Table II**.

As expected, the three known antibiotic gene clusters of *S. clavuligerus* were identified in our genome assembly. Although the supercluster encoding the clavulanic acid and cephamycin C biosynthetic pathways (SMC10-11) and one of the clavam clusters (Tahlan et al. 2007) (SMC9) lie on the main chromosome, the alanylclavam cluster (SMCp13) (Zelyas et al. 2008) is located on the megaplasmid. The latter cluster has been called a 'paralogous cluster' of the clavulanic acid gene cluster (Tahlan et al. 2004), and was thought to have arisen through a recent gene duplication. Interestingly, a previously unknown third cluster of genes (SMCp25) containing a more distant homolog of the clavaminate synthase gene was also identified on pSCL4. This cluster might encode the production of yet another β-lactam antibiotic.

We could not predict with certainty which gene clusters are responsible for the production of holomycin and the tunicamycin-related antibiotic, given the paucity of knowledge on their biosynthesis. For the tacrolimus-like macrolide, we identified a potential gene cluster: the only macrolide PKS cluster detected, consisting of 11 modules. It is positioned near the end of one of the chromosomal arms (SMC1).

Interestingly, we also identified gene clusters potentially encoding the biosynthesis of known antibiotics on the megaplasmid. Clusters closely homologous to the staurosporine biosynthetic gene cluster of *Streptomyces sp.* TP-A0274 (Onaka et al. 2002) (SMCp14) and the moenomycin biosynthetic gene cluster from *Streptomyces ghanaensis* (Ostash, Saghatelian, Walker 2007; Ostash et al. 2009) (SMCp18) were detected, as well as a close homolog of the IndC indigoidine blue pigment synthetase from *Streptomyces lavendulae* (Takahashi et al. 2007) (SMCp24). The fact that pSCL4 carries multiple biosynthetic gene clusters closely resembling those in relatively distantly related *Streptomyces* species, while the same clusters are absent in more closely related species (**Supplementary Figure 6**), supports the hypothesis that many secondary metabolite biosynthetic gene clusters in bacteria are acquired by horizontal gene transfer.

Putative gene clusters that might encode unknown products were also identified on both the megaplasmid and the chromosome. These include two enediyne-type PKS clusters on the megaplasmid (SMCp16 and SMCp21), which have the typical *unbLVU* genes encoding the biosynthetic machinery for the core enediyne structure. The clusters are quite similar to the biosynthetic gene clusters of C-1027 (Liu et al. 2002) and calicheamycin (Ahlert et al. 2002). Both *S. clavuligerus* clusters fall into the phylogenetic subgroup of 9-membered enediyne polyketides that

was recently identified (Udwary et al. 2007) (**Supplementary Figure 7**), although we should note that recombination might have occurred between the two clusters (we observed some extremely similar regions in the multiple sequence alignments of these PKSs) which might distort the phylogenetic analysis.

The second enediyne-type PKS cluster lies very close to an NRPS gene cluster (SMCp20) with a very unusual feature that has not been observed before: the module containing the thioesterase is fused C-terminally to a major facilitator-type transporter. This may be used to export the end-product from the cell immediately after assembly. Such a mechanism could have the advantage that the transport efficiency of the substance is much higher, and the cell can synthesize highly toxic compounds without these poisoning the producing cell.

Yet another remarkable cluster (SMC14) on the chromosome has one of its putative NRPS modules fused to a beta-lactamase domain (PFAM number PF00144). No significant resemblance of this domain to any experimentally studied protein was detected, so we can only speculate about its function: it might for example act as a transpeptidase or bind to, but not hydrolyze, a beta-lactam compound that is then attached to the peptide synthesized by the NRPS. Generally, these domains are of high importance in beta-lactam producing organisms, as they can catalyse the opening and hydrolysis of the beta-lactam ring of beta-lactam antibiotics, and can thus provide resistance to the strain's own antibiotics. Twenty-two proteins carrying a predicted beta-lactamase domain were detected on the chromosome or plasmid (**Supplementary Table III**).

Cross-regulation takes place between the megaplasmid and the chromosome

Expression of secondary metabolite genes can be regulated in a variety of ways. One important regulatory mechanism involves small signalling molecules called γ-butyrolactones (Takano 2006). In *S. clavuligerus*, a gene encoding a γ-butyrolactone receptor protein (ScaR / Brp) was recently identified and shown to regulate clavulanic acid and cephamycin C production (Kim et al. 2004; Santamarta et al. 2005). We identified this gene as the megaplasmid-encoded SCLAV_p0894, which appeared to be the only γ-butyrolactone receptor protein gene in the entire genome. However, four ScbA/AfsA-like putative butyrolactone biosynthetic proteins were detected: three on the chromosome (SCLAV_0463, SCLAV_0471, SCLAV_2310) and one (SCLAV_p0812) on the megaplasmid. Based on phylogenetic analysis (**Supplementary Figure 8**), SCLAV_2310 is the most likely candidate for being involved in butyrolactone biosynthesis. The fact that the only γ-butyrolactone receptor protein is encoded on the megaplasmid is remarkable, as all other characterized γ-butyrolactone receptors are chromosomally encoded. Moreover, it means that Brp trans-regulates several factors on the chromosome (at least the cephamycin C and clavulanic acid gene clusters). Because in a phylogenetic analysis Brp was shown to cluster together with all known chromosomally encoded γ-butyrolactone receptors (Nishida et al. 2007), it is tempting to speculate that this gene was originally derived from the main chromosome and became plasmid-borne through recombination or transposition. Interestingly, two of the AfsA domain-containing butyrolactone biosynthetic proteins (SCLAV_0463 and SCLAV_0471) lie in a PKS gene cluster (SMC5), which consequently might be regulated by butyrolactone signalling.

Other regulatory genes were also identified. Fifty sigma factor genes were found, of which 43 lie on the chromosome and seven on pSCL4. Fifty-one paired two-component regulatory systems were identified, of which 44 are encoded on the chromosome and seven on the megaplasmid. Also, 21 orphan response regulator genes and eight orphan histidine kinase genes were identified. The two-component systems include at least five (chromosomally encoded) very close homologues of systems that have been observed to be involved in the regulation of antibiotic biosynthesis: AfsQ12 (SCLAV_3812–3813; Ishizuka et al. 1992), CutRS (SCLAV_4778–4779; Chang et al. 1996), RapA12 (SCLAV_4312–4313 ; Lu et al. 2007), AbrB12 (SCLAV_1381–1382 ; Yepes et al. 2009) and AbiA123 (SCLAV_3595–3597 ; Yepes et al. 2009). Additionally, 26 serine/threonine kinase genes were detected, of which only two are on the megaplasmid. Finally, 20 genes encoding *Streptomyces* antibiotic regulatory protein (SARPs) were detected, of which eight are on the megaplasmid. Seven SARP-encoding genes lie in or very near secondary metabolism gene clusters.

The vast array of potential regulatory mechanisms suggests that the production of secondary metabolites in *S. clavuligerus* — as in other streptomycetes — is under complex regulation, being highly tuned to the specific needs of the organism. It is striking that for all classes of regulators the plasmid seems to encode considerably fewer compared to the chromosome, reinforcing the view that it is a highly specialized genetic element.

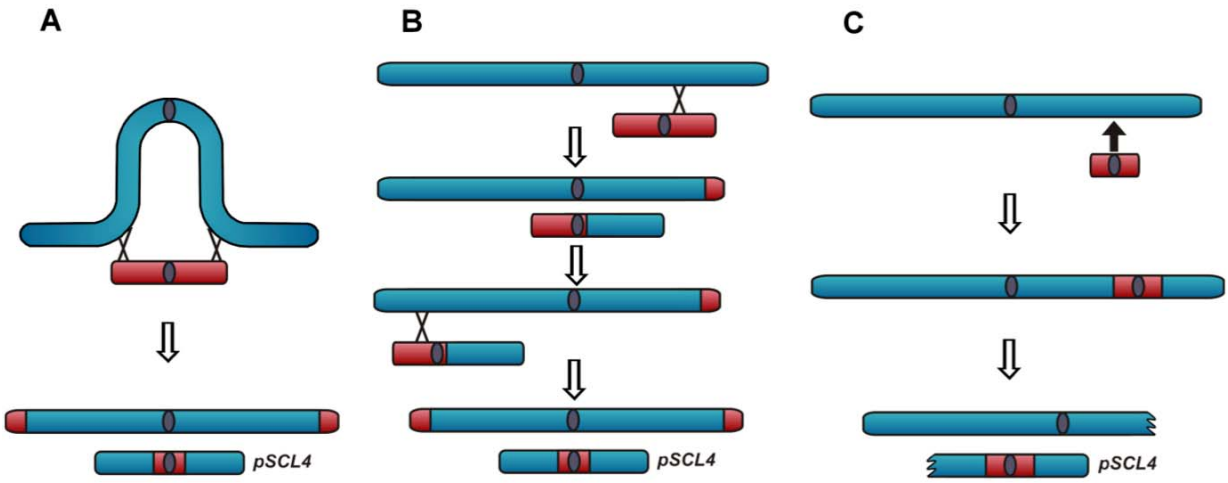

Figure 3: Three scenarios for the evolution of megaplasmid pSCL4. A) A double crossover between the chromosomal core and the core of a smaller plasmid. B) Two consecutive recombination events, the first yielding an asymmetrical plasmid, the second a symmetrical one. C) Integration of a relatively small plasmid into a chromosomal arm and subsequent breaking off of the arm to yield an independent replicon.

Scenarios for the evolution of pSCL4

The small size of the *S. clavuligerus* chromosome and the functional similarity between pSCL4 and *Streptomyces* chromosomal arms prompted the question how a megaplasmid like pSCL4 could have originated. Interestingly, a megaplasmid or chromosome of very similar size (1.8 Mb) was observed in *S. coelicolor* A3(2), after a single crossover between the 365-kb SCP1 plasmid and the chromosome (Yamasaki and Kinashi 2004). However, such a simple single crossover cannot have produced a plasmid configuration as seen in pSCL4, as it leads to an asymmetrical replicon with the

Ori distant from the centre (**Figure 3B**, upper part). The central position of the pSCL4 Ori suggests that multiple recombination events have taken place, either simultaneously or consecutively. Furthermore, the small size of the main chromosome compared to those of other *Streptomyces* species suggests that the plasmid may well have been endowed with genetic regions from the chromosome arms in the recent past. Other observations that should also be taken into account are the clustering of replication genes (*parAB*, *tap*/*tpg*, plasmid primase/replicase genes) and other important genes such as the butyrolactone receptor gene *brp* close to the plasmid Ori, and the existence of a region close to the plasmid Ori (SCLAV_p0926–0939) similar to chromosomal regions of many other actinomycetes, but not found on the *S. clavuligerus* chromosome. These considerations suggest three possible scenarios for the origin of pSCL4 (**Figure 3**).

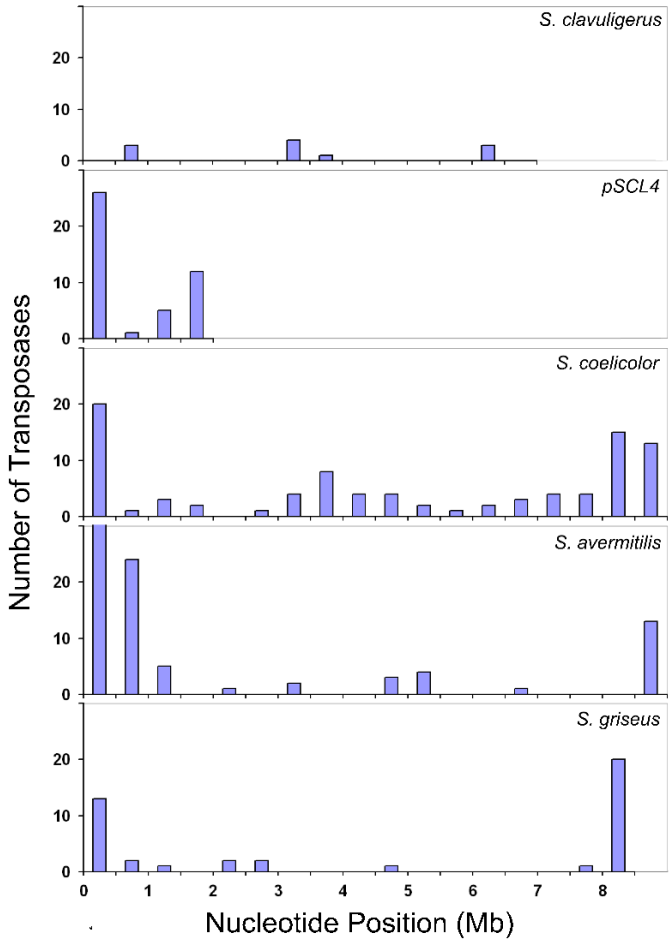

Figure 4: Transposases in *Streptomyces* genomes. Transposases identified in the *S. clavuligerus* genome, compared to the transposases identified in the genomes of *S. coelicolor*, *S. avermitilis*, and *S. griseus*. In contrast to the three other genomes, the chromosome of *S. clavuligerus* does not encode a large number of transposases at its ends. However, pSCL4 does encode many transposases. The fact that this feature – typical of chromosomal ends – is not present in the chromosome yet present in the plasmid supports the hypothesis that the plasmid originated by acquiring the terminal regions of the chromosome.

According to the first scenario, a relatively small plasmid underwent double crossing over with the chromosomal core, to yield a core chromosome with small plasmid arms and a plasmid core with large chromosomal arms. The second scenario is a variant of the first, but postulates two asynchronous recombination events: first a smaller plasmid recombined with one chromosomal arm, yielding an asymmetrical plasmid, and then this situation was stabilized by another recombination

with the other chromosomal arm. In the third scenario, a small integrative plasmid would have integrated into one of the chromosomal arms, after which it would have broken off (e.g. during conjugative transfer) and become a separate replicon.

The first two scenarios both seem parsimonious, as they can explain the symmetrical GC skew plot of pSCL4 without the need for extensive subsequent evolutionary fine-tuning. Moreover, these scenarios are also favored over the third because the main chromosome is quite symmetrical despite its small size, and has long arms on neither side. A final argument supporting the first and second scenarios is provided by the distribution of transposons throughout the *S. clavuligerus* genome. Normally, many transposons are present at the far termini of the main chromosomes of streptomycetes (Chen et al. 2002), yet they are absent from the termini of the *S. clavuligerus* chromosome. Instead, many transposons are present near the termini of pSCL4 (44 in total, judged by the number of transposases found; **Figure 4**). Intriguingly, phylogenetic analysis showed that most of these (25 of the 44) are close homologues of transposases encoded in *Streptomyces* chromosomal termini (<1 Mb distance to the ends). This suggests that the plasmid may indeed contain the former chromosomal termini of *S. clavuligerus*. Sequencing of the telomeres at the very ends of chromosome and megaplasmid might help to confirm this.

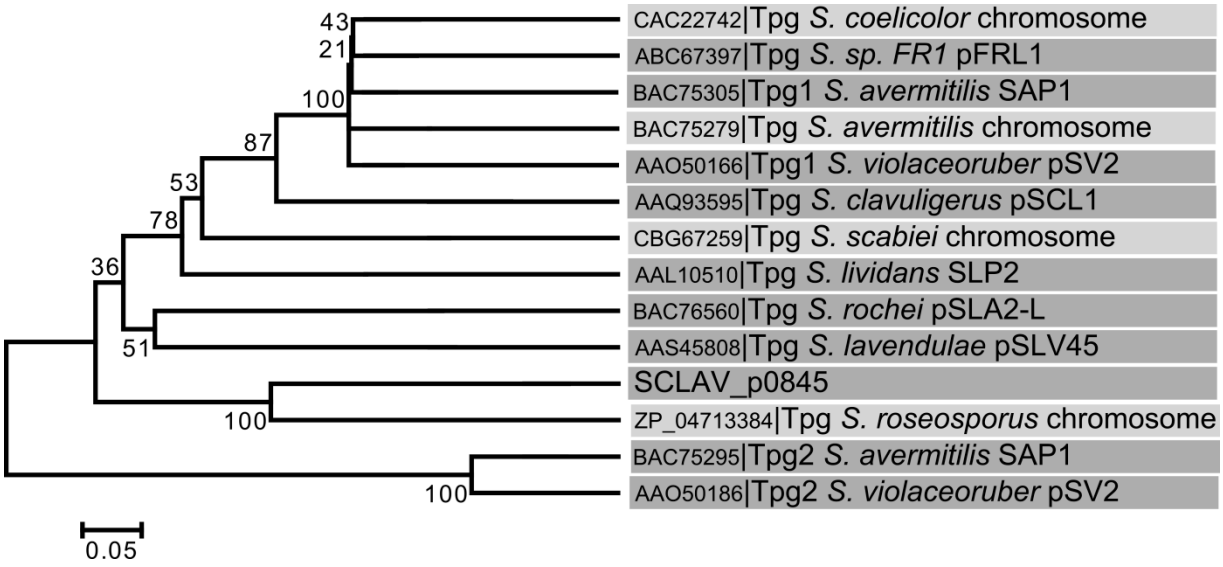

Figure 5: Phylogeny of TPG proteins shows the dynamic evolution of linear plasmids and chromosomes in *Streptomyces*. Phylogenetic analysis of Tpg proteins shows that there is no correlation between the phylogenetic clustering and presence of a *tpg* gene on either the plasmid (in dark gray) or the chromosome (in light gray). This suggests that *tap* and *tpg* genes are often transferred from chromosomes to plasmids and *vice versa*, e.g. through recombination. Phylogenies were calculated using the NJ method in MEGA 4 (Tamura et al. 2007). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) is shown next to the branches.

Heavy traffic between plasmids and chromosomal arms in actinomycete bacteria

Our findings underline the dynamic evolution of actinomycete genomes. They seem to acquire or lose plasmids regularly, and the gene content of their chromosomal arms fluctuates dramatically (Choulet et al. 2006). This is clearly visible in the phylogeny of the Tpg telomere proteins (**Figure 5**), which shows almost complete intermixing of chromosome- and plasmid-encoded genes, signifying

extensive transfer of *tpg* genes via recombination or integration. Our identification of a very large plasmid densely packed with genetic material that is typically encoded on chromosomal arms suggests that linear plasmids play a major role in this genomic flux. Because of their ability to reach large sizes and the possibility of plasmid-chromosome recombination or plasmid integration (Kinashi, Shimaji-Murayama, Hanafusa 1992) such as described above, plasmids seem to more important than previously thought in determining the large variability of actinomycete genomes. This hypothesis fits neatly with the observation that many actinomycete species-specific secondary metabolite clusters are located within genomic islands (Penn et al. 2009), which are often mobilized on plasmids (Dobrindt et al. 2004). The fact that the *S. clavuligerus* ATCC 27064 strain appears to have a very different genome composition in terms of plasmids than the NRRL 3585 strain (Netolitzky et al. 1995) — which originates from the same source (Higgens and Kastner 1971) — epitomizes the highly dynamic nature of actinomycete plasmids. Earlier observations of rapid recombination events between chromosomes and plasmids (Gravius et al. 1994; McLeod et al. 2006) in actinomycetes with linear chromosomes had already suggested such a conclusion.

Conclusions

The sequencing of the genome of *S. clavuligerus* reveals the potential for producing a vast array of interesting novel secondary metabolites, some of them encoded by unusual gene configurations that have never been observed before. The fact that a major part of the secondary metabolite biosynthesis gene clusters is localized on a specialized giant linear plasmid (pSCL4), which we predict not to contain any genes essential for primary metabolism, indicates that plasmids could more often encode secondary metabolites than previously thought (Kinashi 2008). The small size of the *S. clavuligerus* chromosome and the absence of hallmarks typical for *Streptomyces* chromosomal termini, together with the remarkable presence of such terminal hallmarks on the plasmid, suggest that the megaplasmid may have originated by a double recombination of a smaller plasmid with the chromosome. There is evidence that cross-regulation of chromosomal genes by at least one plasmid-encoded regulator still occurs. Intriguingly, the deduced flow of genetic material between different replicons may not be an exception: the phylogeny of telomere replication proteins from plasmids as well as chromosomes suggests that fluxes of genetic material regularly take place between streptomycete chromosomes and plasmids. Indeed, the mobilization of secondary metabolite gene clusters onto large linear plasmids such as pSCL4, which could be vectors for horizontal gene transfer (Ravel, Schrempf, Hill 1998; Ravel, Wellington, Hill 2000), indicates that constant and extensive 'open source' evolution (Frost et al. 2005) of secondary metabolite-encoding DNA regions in actinomycetes could be responsible for the large differences in secondary metabolite repertoires between different species. Approaches that "awaken" uncharacterized gene clusters will undoubtedly be pivotal to uncover the functionalities of the secondary metabolites they encode (Scherlach and Hertweck 2009). Furthermore, if the plasmid could be cured from the strain (Hsu and Chen 2010), the small chromosome of *S. clavuligerus* may be a very interesting vehicle for synthetic biology (Keasling 2008), serving as a starting point for the construction of a 'minimal streptomycete.'

Acknowledgments

This work was supported by the Dutch Technology Foundation STW, which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs [STW 10463]. RB is supported by an NWO-Vidi fellowship, and ET by a Rosalind Franklin Fellowship, University of Groningen. We thank Christian Kuijlaars for help in the initial phase of metabolic model construction. We thank David Hopwood and the anonymous reviewers for constructive comments and suggestions.

Supplementary Material

Supplementary figures and tables can be downloaded from http://rdmy.info/ch8

Chapter 9

Genome-wide gene expression changes in an industrial clavulanic acid overproduction strain of *Streptomyces clavuligerus*

With addendum: The future of industrial antibiotic production: from random mutagenesis to synthetic biology.

Published as:

- M.H. Medema*, M.T. Alam*, W.H. Heijne, M.A. van den Berg, U. Müller, A. Trefzer, R.A. Bovenberg, R. Breitling, E. Takano (2011) Genome-wide gene expression changes in an industrial clavulanic acid overproduction strain of *Streptomyces clavuligerus*. *Microbial Biotechnology* 4: 300-305.
- M.H. Medema*, M.T. Alam*, R. Breitling, E. Takano (2011) The future of industrial antibiotic production: from random mutagenesis to synthetic biology. *Bioengineered Bugs* 2: 230-233.

*Equal contribution

Abstract

To increase production of the important pharmaceutical compound clavulanic acid, a β-lactamase inhibitor, both random mutagenesis approaches and rational engineering of *Streptomyces clavuligerus* strains have been extensively applied. Here, for the first time, we compared genome-wide gene expression of an industrial *S. clavuligerus* strain obtained through iterative mutagenesis, with that of the wild type strain. Intriguingly, we found that the majority of the changes contributed not to a complex rewiring of primary metabolism but consisted of a simple upregulation of various antibiotic biosynthesis gene clusters. A few additional transcriptional changes in primary metabolism at key points seem to help to divert metabolic fluxes to the biosynthetic precursors for clavulanic acid. In general, the observed changes largely coincide with genes that have been targeted by rational engineering in recent years, yet the upregulation of a number of previously unexplored genes clearly demonstrates that functional genomic analysis can provide new leads for strain improvement in biotechnology.

Introduction

Streptomyces clavuligerus is an important industrial microorganism, which produces the β-lactam antibiotic cephamycin C (Martin and Liras 1989) and the β-lactamase inhibitor clavulanic acid (Saudagar, Survase, Singhal 2008). Clavulanic acid is produced worldwide on a large scale, and co-formulated with amoxicillin in Augmentin® (Brogden et al. 1981). Two biotechnological strain optimization approaches have been utilized to increase the production of clavulanic acid by the bacterium: rational metabolic engineering and iterative optimization through random mutagenesis and screening.

In the rational approach, a specific gene is knocked out or overexpressed to divert metabolic fluxes towards the antibiotic biosynthetic pathways. Arguably the best example comes from the work of (Li and Townsend 2006), who re-engineered the *S. clavuligerus* glycolytic pathway by constructing a deletion mutant of the glyceraldehyde-3-phosphate dehydrogenase gene *gap1* to increase the pool of the clavulanic acid precursor glycerol-3-phosphate (G3P). This doubled clavulanic acid production compared to the wild type. Overexpression of the regulatory proteins CcaR and ClaR also led to clavulanic acid overproduction (Hung et al. 2007), and the two strategies have recently been combined successfully in a single strain (Jnawali, Lee, Sohng 2010).

Yet, most or even all production strains that are used in industry have been obtained by classical strain improvement (Adrio and Demain 2006), based on mutagenesis with mutagens such as nitrosoguanidine (NTG). Little is known about the exact genetic changes through which high production titers are achieved in these mutants.

Recently, we published the genome sequence of *S. clavuligerus* ATCC 27064 (Medema et al. 2010). We now employed this information to perform a genome-wide transcriptome study on an industrial production strain which has been generated from the ATCC 27064 type strain (Higgens and Kastner 1971) by several iterations of mutagenesis and screening, and produces clavulanic acid at levels approximately 100x that of the wildtype. The majority of observed changes consists of increased transcript levels of the antibiotic biosynthesis gene clusters. These observations are in agreement

with flux-balance analysis predictions using a constraint-based genome-scale metabolic model constructed for *S. clavuligerus*. However, we also detected some potentially crucial transcript level changes in primary metabolism that could contribute to the increased production of clavulanic acid by redirection of fluxes, mimicking strategies utilized in rational approaches.

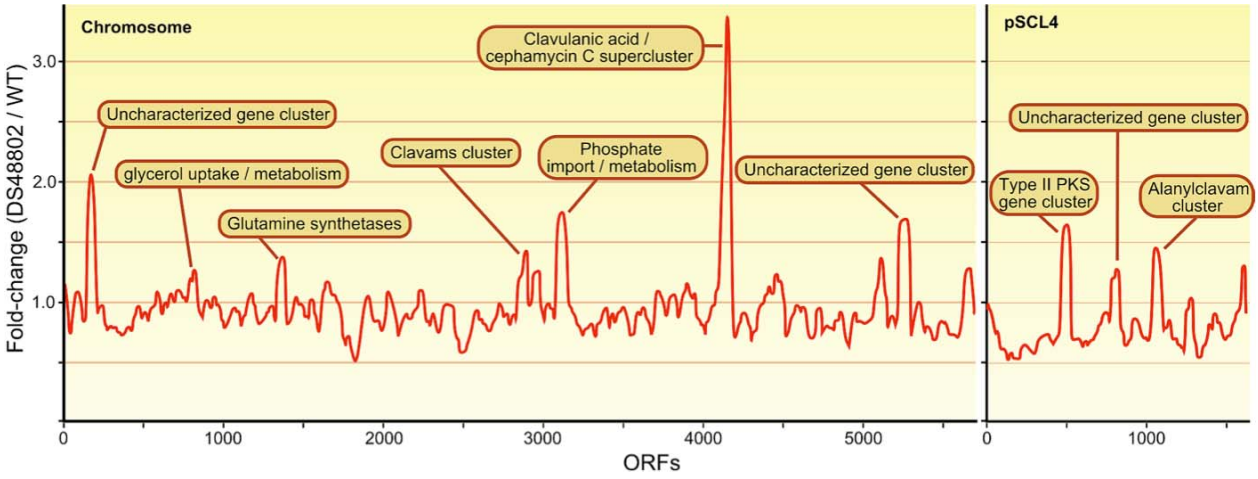

Figure 1: Differential gene expression in *S. clavuligerus* DS48802 and ATCC 27064. Sliding window plot (size = 50) of the difference in gene expression between *S. clavuligerus* DS48802 and wild type ATCC 27064. Key upregulated operons or genes at the peaks are noted in the figure. See **Supplementary Table II** for description of the ten gene clusters shown. For gene expression analysis, cultivations were performed in shake flasks directly inoculated with spore suspensions at 28ºC and 280 rpm. The semi-synthetic growth medium used consisted of 30 g/ l glycerol, 5 g/l wheat gluten, 3.5 g/l asparagine monohydrate, 1.5 g/l L-lysine, 0.7 g/l KH_2PO_4, 0.3 g/l $MgSO_4 \cdot 7H_2O$, 0.2 g/l $CaCl_2 \cdot 2H_2O$, 0.2 g/l $FeSO_4 \cdot 7H_2O$, 10 g/l MOPS, 0.1 ml/l Basilodon, and 1 ml/l trace elements solution at pH7.0. The trace element solution consists of 20.4 g/l H_2SO_4, 50 g/l citric acid H_2O , 16.75 g/l $ZnSO_4 \cdot 7H_2O$, 1.6 g/l $CuSO_4$. 5 H_2O, 1.5 g/l $MnCl_2 \cdot 4H_2O$, 2 g/l H_3BO_3, 2 g/l $Na_2MoO_4 \cdot 2H_2O$. After 70 h of cultivation, the cells were harvested by centrifugation, treated with RNAprotect (Qiagen) and directly frozen with liquid nitrogen and stored at −80°C. To isolate total RNA, the frozen mycelium was ground in a mortar, re-suspended in TE buffer with 5 mg/l lysozyme and incubated for 5 min at room temperature. RNA isolation and purification were performed using phenol extraction (TRIzol reagent, Invitrogen) and RNeasy Kit (Qiagen). The RNA was quantified by measuring the absorbance at 260 nm. Biotinylated cDNA was prepared after fragmentation according to the standard Affymetrix protocol using GC rich (average 72%) primers from 10 µg total RNA. For hybridization, 5 µg and 7 µg biotinylated cDNA were used per Affymetrix gene Chip. Microarray data have been deposited at Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession number *GSE24033*. Flux-balance analysis was performed using a recently published genome-scale metabolic model of *S. clavuligerus* (Medema et al., 2010). In this study, we slightly changed our objective function and included both clavulanic acid and cephamycin C biosynthesis pathways. We dynamically changed the antibiotic concentration in the biomass composition based on experimental observations of Romero et al. (1985) and optimized the objective function for different concentrations of each antibiotic. Among the 785 genes that the model contains, 497 genes showed non-zero flux for at least one antibiotic concentration. We calculated Spearman correlation of fluxes of each reaction with increasing antibiotic concentrations. If an enzyme was involved in multiple reactions, we assigned the flux that had the highest r^2.

Results and Discussion

Increased transcription of secondary metabolite biosynthesis gene clusters in strain DS48802

When we compared gene transcript levels of *S. clavuligerus* wild type and DS48802 strains during stationary phase using microarrays, almost all genes ranking high in a differential transcriptome

analysis appeared to belong to the complete clavulanic acid / cephamycin C supercluster, which is significantly overexpressed (between two- and eight-fold) in the DS48802 strain compared to the wild type. Interestingly, the pathway-specific regulator genes *claR* and *ccaR* are also overexpressed in DS48802. They are located within the same supercluster and their products have been shown to regulate it positively (Alexander and Jensen 1998; Paradkar, Aidoo, Jensen 1998).

Additionally, the clavams gene cluster (*cvm1245* and *cas1*) and the 'paralogous' alanylclavam cluster (*orfABCD* and *ceaS1*/*pah1*/*bls1*/*oat1*) are significantly overexpressed (**Figure 1**), as is the two-component system involving Cvm7p (SCLAV_p1079-p1080) that induces expression of the alanylclavam cluster (Tahlan et al. 2007). This suggests the presence of a regulatory mechanism common to all these clusters. CcaR is an unlikely candidate for such a common regulatory factor, as it does not appear to control the paralogous cluster (Tahlan, Anders, Jensen 2004); the pleiotropic regulator AdpA (SCLAV_1957) is a more likely candidate, as it is known to induce clavulanic acid expression (Lopez-Garcia, Santamarta, Liras 2010) and its gene is transcribed almost 2.5 times stronger in DS48802.

In contrast to, for example, the intrachromosomal amplification of the kanamycin biosynthesis gene cluster in *Streptomyces kanamyceticus* (Yanai, Murakami, Bibb 2006), hybridization of *S. clavuligerus* DS48802 genomic DNA to the microarrays revealed no amplifications of genes or gene clusters (data not shown). The overproduction that we observe therefore appears to be caused by transcriptional (and post-transcriptional) changes only.

Flux-balance analysis of increased clavulanic acid production correlates well with transcriptomic data

While many changes have occurred during the generation of the DS48802 strain through mutagenesis, it is important to note that changes in transcription levels of many genes may be due to random mutations that have no impact on antibiotic biosynthesis. In order to assess which changes could be causatively linked to antibiotic overproduction, we computationally predicted the metabolic fluxes during antibiotic overproduction, using a constraints-based genome-scale metabolic network model of *S. clavuligerus* (Medema et al. 2010). We dynamically modelled the metabolic flux changes during increased production of clavulanic acid and cephamycin C with different rates of antibiotic production relative to the biomass, based on the experimental observations of (Romero, Liras, Martin 1986). Interestingly, the computational predictions made through dynamic flux-balance analysis (FBA) are largely in line with the observed expression changes. Eighty-seven genes were predicted to be upregulated (positively correlated with antibiotic production; $r > 0.6$) and 129 genes were predicted to be down-regulated (negatively correlated; $r < -0.6$) at increasing antibiotic production levels. Forty percent (15/37) of the genes that actually showed increased transcript levels (fold-change > 2) were also predicted to do so according to FBA, and these include all genes encoding key biosynthetic enzymes known to be involved in clavulanic acid and cephamycin C biosynthesis. Even though 40% does not appear to be a very large percentage, these predictions are clearly statistically significant according to a Fisher's exact test (p-value 0.0005; see **Supplementary Table I**). One should note that FBA predicts the flux for every reaction, not for every gene product, as multiple gene products can be involved in a single reaction. Therefore, the same flux was assigned to all gene products involved in that particular reaction, which is not necessarily the case in the actual gene expression, as a single homologue or isoenzyme could

be actively performing the reaction and thus would be differentially expressed. Out of the 47 different enzymatic reactions predicted to be upregulated (associated with 87 genes), 26 (55%) have at least one gene linked to them which showed increased transcript levels. As both our FBA and gene transcript analysis pointed to an increased expression of the clavulanic acid and cephamycin core biosynthesis genes, we suggest that this is a crucial change required for antibiotic overproduction in this strain. Moreover, as the FBA indicated that the absolute fluxes required for high production of the secondary metabolites are minor compared to other fluxes involved in maintenance and cellular growth, a complete redirection of primary metabolism appears not to be necessary for overproduction.

Gene expression changes in primary metabolism

Nonetheless, because clavulanic acid is synthesized from the precursors G3P and L-arginine, which play important roles in primary metabolism, specific changes in the primary metabolism of DS48802 could have occurred during the various random mutagenesis rounds, so that the intracellular pools of these intermediates are increased.

Indeed, glycerol uptake and metabolism (SCLAV_0631-0632 & SCLAV_0877-0879) is clearly upregulated over twofold in DS48802, indicating an improved utilization of glycerol as a carbon source as well as increased production of the clavulanic acid precursor G3P (**Figure 2**). Moreover, the aconitase and citrate synthase from the citric acid cycle appear to be downregulated. A likely explanation for this is that the carbon flux from G3P in this direction is reduced and is partly redirected to clavulanic acid biosynthesis. This situation is remarkably similar to the result of the rationally constructed *gap1* deletion that blocked G3P conversion into 1,3-bisphosphoglycerate, thus improving clavulanic acid biosynthesis by increasing the intracellular G3P pool (Li and Townsend 2006). However, an advantage of the situation in DS48802, which seems to have an incomplete downregulation of the flux, could be that a considerable pool of acetyl-CoA is maintained, e.g. for the biosynthesis of ornithine from glutamate. DS48802 also seems to avoid the potential negative effects of a complete deletion of the aconitase and citrate synthase genes: a complete absence of these enzyme activities could lead to acidogenesis with negative consequences for secondary metabolite production as shown by (Viollier et al. 2001a; Viollier et al. 2001b). Also, DS48802 still seems to be able to synthesize α-ketoglutarate (a co-substrate required for clavaminic acid biosynthesis; (Salowe, Marsh, Townsend 1990)), while achieving the benefits of higher acetyl-CoA and/or G3P pools that have made these genes attractive targets for rational engineering to improve antibiotic production (Viollier et al. 2001a). A potentially important observation which we cannot explain yet from the current data presents itself in the differential transcript level changes of the two pyruvate kinase isoenzyme genes, of which one is downregulated (SCLAV_4329) and the other upregulated (SCLAV_1203).

We also observed a remarkable upregulation of glutamine synthetases I and II (SCLAV_1416 & SCLAV_1431), glutamate synthetase (SCLAV_1231) and glutamate importers (SCLAV_4660-4663). Glutamate can serve as a source for biosynthesis of the clavulanic acid precursor arginine. This conversion takes place through the urea cycle involving ornithine as an intermediate (Rodríguez-García et al. 2000) addition of which to the medium has been shown to strongly enhance clavulanic acid biosynthesis (Cheng et al. 2003). Probably to overcome nitrogen and phosphate limitations,

genes encoding the transporters for ammonia (SCLAV_4534) and phosphate (SCLAV_3166-3169) are also observed to be more highly transcribed in DS48802. This may be caused by the increased expression of the pathway-specific activator genes *phoU* (SCLAV_3220, (Ghorbel et al. 2006) and *glnB* (SCLAV_4535,(Drepper et al. 2003), which both show over twofold increased transcription.

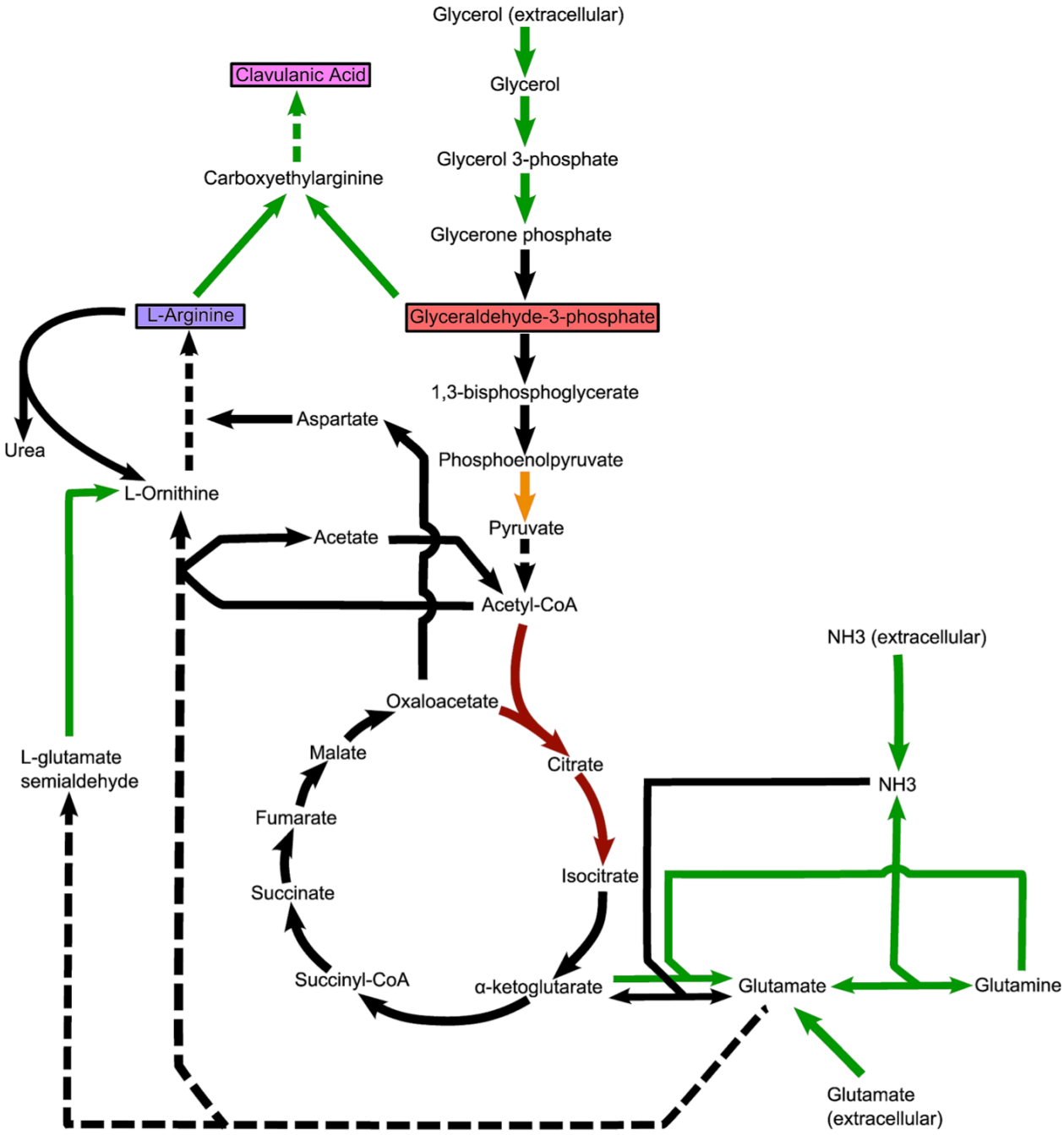

Figure 2: Changes in *S. clavuligerus* primary and secondary metabolism affecting clavulanic acid production. Changes in gene expression in *S. clavuligerus* DS48802 compared to the wild type ATCC 27064 projected onto a metabolic map. Green arrows represent reactions catalyzed by genes expressed over twofold higher in DS48802 than in the wild type. Red arrows represent reactions catalyzed by genes expressed over two-fold lower in DS48802. The orange arrow represents the reaction catalyzed by pyruvate kinase, for which two isoenzymes exist which have changed in expression differently, one being downregulated (SCLAV_4329) and the other being upregulated (SCLAV_1203). Black arrows represent unchanged steps. Solid arrows represent single biosynthetic steps, dashed arrows represent multiple steps.

Conclusions

Our data show that a strain improvement program by random mutagenesis and screening has caused gene transcript changes in both primary and secondary metabolism. The overlap with results obtained by rational metabolic engineering through *claR*/*ccaR* overexpression and *gap1* deletion is intriguing. New leads from transcript changes observed in this study, such as the increased transcription of glutamine and glutamate synthetase genes, and of those encoding ammonium and phosphate transporters, could be combined to rationally design novel high-producer strains. This approach might avoid the introduction of unwanted adverse effects from random mutagenesis, and provide strains suited for industrial application in a more efficient way. In this manner, functional genomics allows two key strategies applied in biotechnology — random mutagenesis and rational engineering — to become increasingly complementary.

Acknowledgements

This work was supported by the Dutch Technology Foundation STW, which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs [STW 10463]. RBr is supported by an NWO-Vidi fellowship, and ET by a Rosalind Franklin Fellowship, University of Groningen. We thank Bert Koekman for providing the classically improved strain DS48802. We thank Hilde Huininga and Hildegard Menke for technical assistance with DNA and RNA isolations.

Supplementary Material

Supplementary tables can be downloaded from http://rdmy.info/ch9

Addendum: The future of industrial antibiotic production: from random mutagenesis to synthetic biology

Natural products derived from the secondary metabolism of microbes constitute a cornerstone of modern medicine. Engineering bugs to produce these products in high quantities is a major challenge for biotechnology, which has usually been tackled by either one of two strategies: iterative random mutagenesis or rational design. Recently, we analyzed the transcriptome of a *Streptomyces clavuligerus* strain optimized for production of the β-lactamase inhibitor clavulanic acid by multiple rounds of mutagenesis and selection, and discovered that the observed changes matched surprisingly well with simple changes that have been introduced into these strains by rational engineering. Here, we discuss how in the new field of synthetic biology, random mutagenesis and rational engineering can be implemented complementarily in ways which may enable one to go beyond the status quo that has now been reached by each method independently.

In our recent paper in *Microbial Biotechnology* (Medema et al. 2011c), we compared levels of gene expression between a *S. clavuligerus* clavulanic acid overproduction mutant (DS48802) and the wildtype. The changes caused by random mutagenesis were strikingly similar to those rationally engineered by two strategies that have recently been employed to increase clavulanic acid production in *S. clavuligerus*: redirection of carbon fluxes towards the key clavulanic acid precursor glyceraldehyde-3-phosphate (G3P) (Li and Townsend 2006) and upregulation of pathway-specific activators (Hung et al. 2007; Jnawali, Lee, Sohng 2010). More precisely, we found the pathway-specific activator genes *ccaR* and *claR* to be more highly expressed in strain DS48802, and we also observed changes in transcript levels of genes associated with glycolysis and the citric acid cycle, which appear to match the intended effect of the recently engineered *gap1* deletion mutant, leading to a redirection of carbon fluxes towards G3P. A third way in which overproduction of chemicals has been achieved by rational engineering is the genomic duplication or even amplification of the biosynthetic gene cluster, such has recently been successfully implemented for overproduction of platensimycin (Smanski et al. 2009) and nikkomycin (Liao et al. 2010). Although a similar amplification has been observed previously in a randomly mutagenized kanamycin overproduction strain (Yanai, Murakami, Bibb 2006), we did not observe any amplification of the clavulanic acid biosynthetic gene cluster on genomic DNA hybridizations to our microarrays. However, the observed overexpression of the clavams biosynthetic gene clusters — which are strongly related to the clavulanic acid gene cluster — in strain DS48802 may also point to recruitment of enzymes for clavulanic acid overproduction from these homologous pathways, alleviating the need for amplification of the original gene cluster. Even though in this case it would be an enigma why cvm5 and cvm6p, which are thought to be specifically involved in the final steps of 5S clavam biosynthesis (Tahlan et al. 2007; Zelyas et al. 2008), are also overexpressed in DS48802, the general picture shows that random mutagenesis and traditional engineering yield similar results: small adjustments to the metabolic or regulatory network of the cell which allow finding a local optimum of production given the rest of the network.

As the field of synthetic biology is maturing, the methodologies for rationally engineering bacterial strains for production of natural products are drastically changing (Medema et al. 2011a). With these

changes come new prospects to make random mutagenesis approaches complement rational engineering in novel ways, in a 'next generation' synthetic biology approach. Synthetic biology engineering is by definition not restricted to the natural architecture of a certain bacterial strain: cellular systems can be extensively re-engineered, or even engineered from the ground up in a *de novo* fashion. But while the general biological knowledge necessary for major re-engineering is often available, the fine-tuning of a design up to the nucleotide level is often much more difficult. This is where random mutagenesis can come in, again to find a local optimum, but this time an optimum which may be much closer to the global optimum because the rest of the metabolic and regulatory network has first been radically modified.

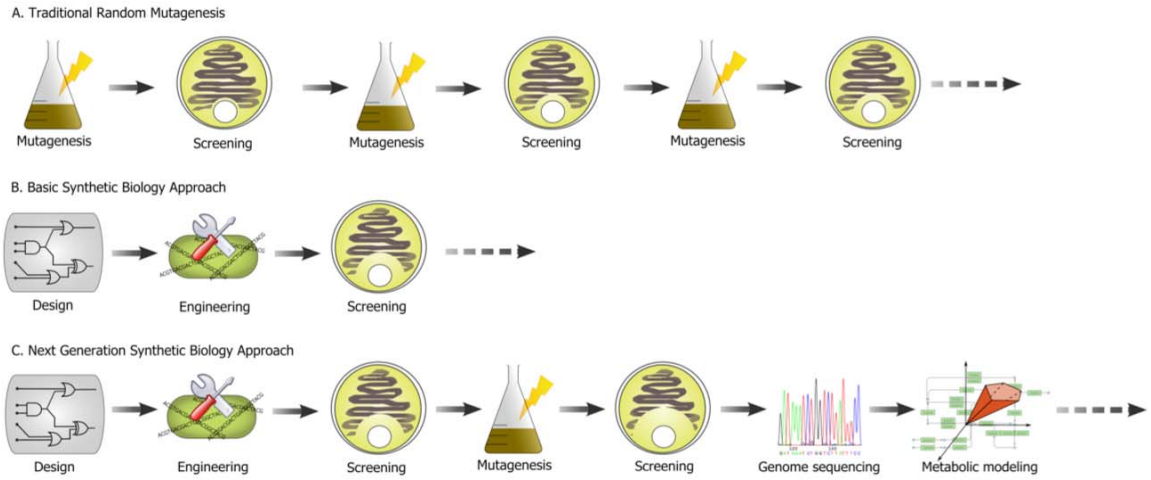

Figure 1. Comparison of several approaches for strain optimization. A. In traditional random mutagenesis, the strain is subjected to several iterations of mutagenesis and screening, in an attempt to arrive at a higher-producing phenotype. B. In the basic synthetic biology approach, an advantage is gained by intelligently engineering a high-producer strain. The strain is designed, engineered, and an activity screen is performed to test the obtained production titers. C. Proposed pipeline for compound production optimization based on a synergy of synthetic biology, systems biology and random mutagenesis. After initial engineering of a first design and screening of its productivity, a round of random mutagenesis of the resulting "designer bug version 1.0" follows. The mutants are screened, and promising improved mutants are selected for genome sequencing. The identified mutations are tested, when possible, with genome-scale metabolic modeling, and combined in engineering a second version of the designer bug. Several additional iterations of design updates or 'bug fixes', mutagenesis and genome sequencing may follow to further optimize the production titers.

As genome sequencing is becoming cheaper by the day, an interesting possibility would be to expose the first version of a design for an overproduction strain to one or a few rounds of mutagenesis, and subsequently sequence the genomes and transcriptomes of a range of promising mutants. In this way, mutations that can help optimize the first design may be identified and combined together without the need for long and tedious repetitions of such rounds of mutagenesis. Metabolic modeling can sometimes aid in the selection of the mutations from the modified genomes, by predicting whether the changes will lead to higher production of the compound. And conversely, the observed profile of mutations can pinpoint bottleneck reactions that may have been missed in the earlier model-driven engineering (**Figure 1**).

When envisaging a *Streptomyces* host for synthetic biology, many efforts have been focussed on generating a minimal *Streptomyces* genome. For example, Komatsu et al. (2010) recently engineered a genome-minimized genome of *Streptomyces avermitilis*, by deleting large regions from the

chromosomal termini. Metabolic modeling may also aid in guiding the construction of such a minimal streptomycete, by predicting the minimal set of metabolic genes necessary for production of biomass.

Recently, we used *in silico* knockouts in a *S. clavuligerus* genome-scale metabolic model to predict that the 1.8-Mb linear plasmid that the strain possesses is not required for primary metabolism and could potentially be cured from the strain (Medema et al. 2010). We used the same model — the version published by Medema et al. (2010), available from the authors in SBML format upon request — to interpret the expression changes seen in the optimized strain, and, again using flux balance analysis on *in silico* knockouts in a minimal medium, we have now attained a first approximation of the minimal *S. clavuligerus* genome. As a start, we generated *in silico* knockouts of all 507 unique enzymes (from 1115 EC-annotated enzymes in total) in the *S. clavuligerus* genome-scale metabolic model. For 159 knockouts (31%), the model was unable to simulate biomass production anymore; the 159 enzymes linked to these knockouts were therefore labeled as unconditionally essential. To be able to identify conditionally essential enzymes as well, we subsequently generated a metabolic network (Ma and Zeng 2003) which links enzyme nodes via edges based on shared common metabolites. Highly connected metabolites were removed from the model reactions, as well as reactions which contain one of these highly connected metabolites as their only product or substrate. We then performed double knockouts by removing all possible pairs of directly connected nodes (enzymes) from the network one by one, and checked the cellular growth again by performing flux balance analysis (Becker et al. 2007). Of the 1079 pairs of enzymes, 657 (61%) were predicted to be essential for biomass production. Seventy enzymes which were not essential according to the first analysis appeared to be essential when knocked out as part of a pair, and were therefore labeled as conditionally essential. Subsequently, we performed triple knockouts by removing three directly connected enzymes at a time. Nineteen new conditionally essential enzymes were detected from this round of knockouts. In total, 194 enzymes (38%) were found nonessential after these three rounds of *in silico* knockouts. However, after removing these 194 predicted non-essential enzymes from the genome-scale model, the model of the cell was not viable during flux balance analysis, probably because in a few cases the conditionality exists on an even higher level, e.g. when there exist two entire entirely independent alternative pathways of which at least one is necessary. Therefore, we manually gap-filled this reduced model by re-adding 49 enzymes, which connect essential enzymes with conditionally essential enzymes, from the set of 195 enzymes that were left from the previous analysis. This successfully allowed simulation of biomass production again.

In the end, we identified a set of 145 enzymes (29%, **Supplementary Table I**) which were predicted to be non-essential (**Figure 2**). These enzymes are encoded by 195 chromosomally encoded genes and 20 plasmid-encoded genes which could potentially be deleted from the *S. clavuligerus* genome to minimize it. As the core chromosome of *S. clavuligerus* is already significantly smaller (6.8 Mb) than the recently published genome-minimized *Streptomyces avermitilis* chromosome (7.6 Mb) (Komatsu et al. 2010), the prospect of being able to further minimize the genome is enthralling. It should be kept in mind that these results were obtained from the analysis of only a fraction of the total 7281 annotated genes from the *S. clavuligerus* genome (i.e., those that encode enzymes). Visualization of the predicted essential and nonessential enzymes showed that the nonessential enzymes primarily reside in pathways involved in secondary metabolism, alternative carbon metabolisms and vitamin and cofactor biosynthesis and alternative carbon metabolisms. The latter could turn out to be essential on different minimal media. Potentially, this information could be used

in an iterative strategy to cut away unnecessary chromosomal regions to further minimize the *Streptomyces* genome after the plasmid megaplasmid (Medema et al. 2010) could perhaps be cured out of the strain, or even — in the long run — to synthesize a *Streptomyces* genome *de novo*.

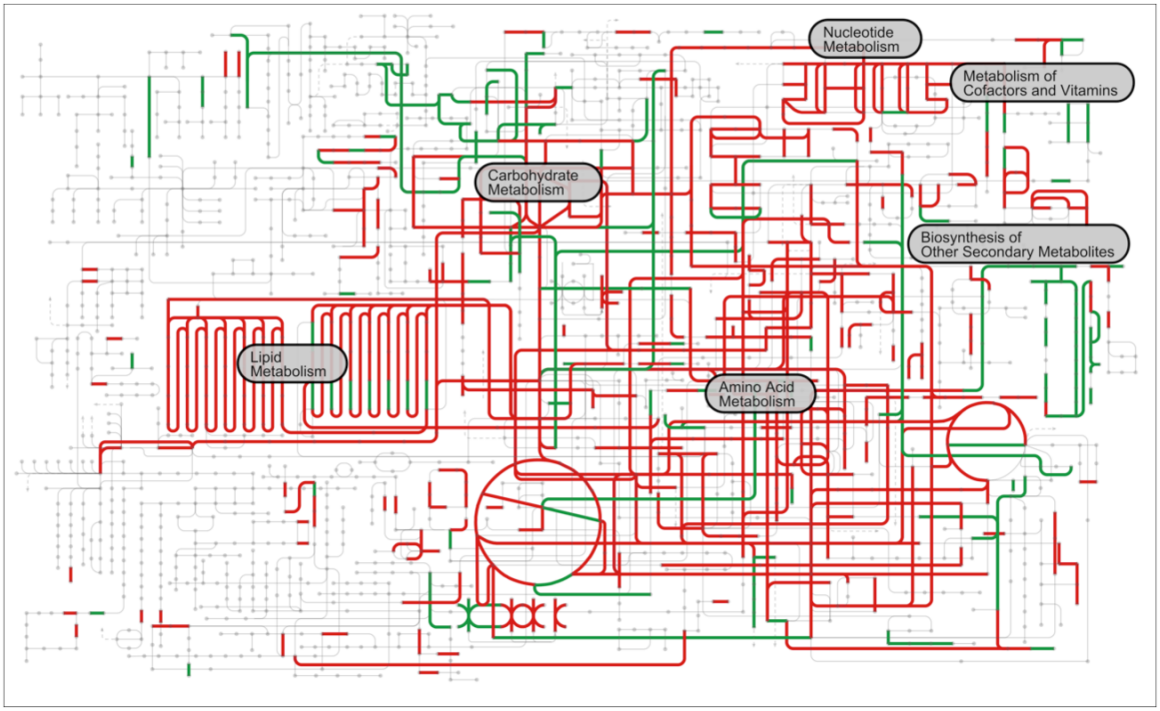

Figure 2. Visualization of essential and nonessential enzymes on the KEGG global pathway map, constructed using the iPATH 2 web server (Letunic et al. 2008). Reactions catalyzed by predicted essential enzymes are colored in red, reactions catalyzed by predicted non-essential enzymes are colored in green.

Overall, although classical strain improvement by random mutagenesis may be loosing its current key position in the generation of overproduction strains, the methodology may still remain valuable for fine-tuning the rough designs produced by synthetic biology. *In silico* modeling of metabolism is likely to complement these approaches by confirming the rationality of mutations, and providing new leads for altering the metabolic network, minimizing the genome and optimizing the whole system for overproduction of the compound of choice.

Chapter 10

Insights into secondary metabolism from a global analysis of biosynthetic gene clusters

Submitted for publication as:

P. Cimermancic*, M.H. Medema*, L.C. Wieland-Brown, K. Mavrommatis, A. Pati, P.A. Godfrey, M. Koehrsen, J. Clardy, B.W. Birren, R. Breitling, E. Takano, A. Sali, M.A. Fischbach (2013) Insights into secondary metabolism from a global analysis of biosynthetic gene clusters. In revision.

*Equal contribution

Abstract

Biosynthetic gene clusters (BGCs) have been discovered for hundreds of bacterial metabolites, including dozens of natural products used in human and veterinary medicine, agriculture, and manufacturing. Although BGCs have been the basis of thousands of genetic and biochemical studies over the last three decades, our knowledge of their number, phylogenetic distribution, and evolution remains limited. Here, we report the results of a systematic effort to identify and analyze BGCs from across the bacterial tree of life. The resulting BGC landscape expands the number of gene cluster classes, reveals the presence of widely distributed gene cluster families of unknown function, and points to numerous unmined organisms with rich biosynthetic potential. The evolution of BGCs is unexpectedly dynamic, with the successive merger of smaller sub-clusters playing a central role in the evolution of larger gene clusters. Moreover, we identify concerted evolution as a key phenomenon that homogenizes domains in a broad range of polyketide synthases and nonribosomal peptide synthetases. Our findings point to novel, widely distributed clusters from undermined taxa, and they suggest novel strategies for biosynthetic gene cluster engineering that mimic Nature's evolutionary process.

Introduction

Connecting small molecules to the genes that encode them is revolutionizing the study of natural products, enabling genome sequence data to guide the discovery of high-value molecules (Bergmann et al. 2007; Challis 2008; Franke, Ishida, Hertweck 2012; Freeman et al. 2012; Kersten et al. 2011; Laureti et al. 2011; Lautru et al. 2005; Letzel, Pidot, Hertweck 2013; Nguyen et al. 2008; Oliynyk et al. 2007; Schneiker et al. 2007; Walsh and Fischbach 2010; Winter, Behnken, Hertweck 2011). The thousands of bacterial genomes in the database provide an opportunity to generalize this approach. Here, we report the results of a systematic effort to identify and analyze biosynthetic gene clusters (BGCs) in 1,154 sequenced genomes spanning the bacterial tree of life. Our analysis focuses on two questions: Where in the landscape of bacterial BGCs do we find the most potential to identify novel natural products? And what are the principles of BGC evolution that could inform future engineering efforts?

Numerous algorithms have been developed for the automated prediction of BGCs in microbial genomes (Khaldi et al. 2010; Li et al. 2009; Medema et al. 2011b; Starcevic et al. 2008; Weber et al. 2009). However, these tools were not suitable for our purposes because they focus exclusively on, or exhibit bias toward, gene clusters for certain classes of molecules (e.g., polyketides or nonribosomal peptides). As a more general solution to the gene cluster identification problem, we developed a hidden Markov model-based algorithm, ClusterFinder, that aims to identify gene clusters of any class (**Figure 1a**, **Methods**, **SI Methods**, and **Supplemental Text 1**). Our method predicted a total of 33,352 putative BCGs with an estimated false-positive rate of 5%, which we divided into two categories – high-confidence (9,421; used in all subsequent analyses, with exceptions noted below) and low-confidence (23,931) – based on whether or not they could be assigned to one of ~20 well-validated BGC classes.

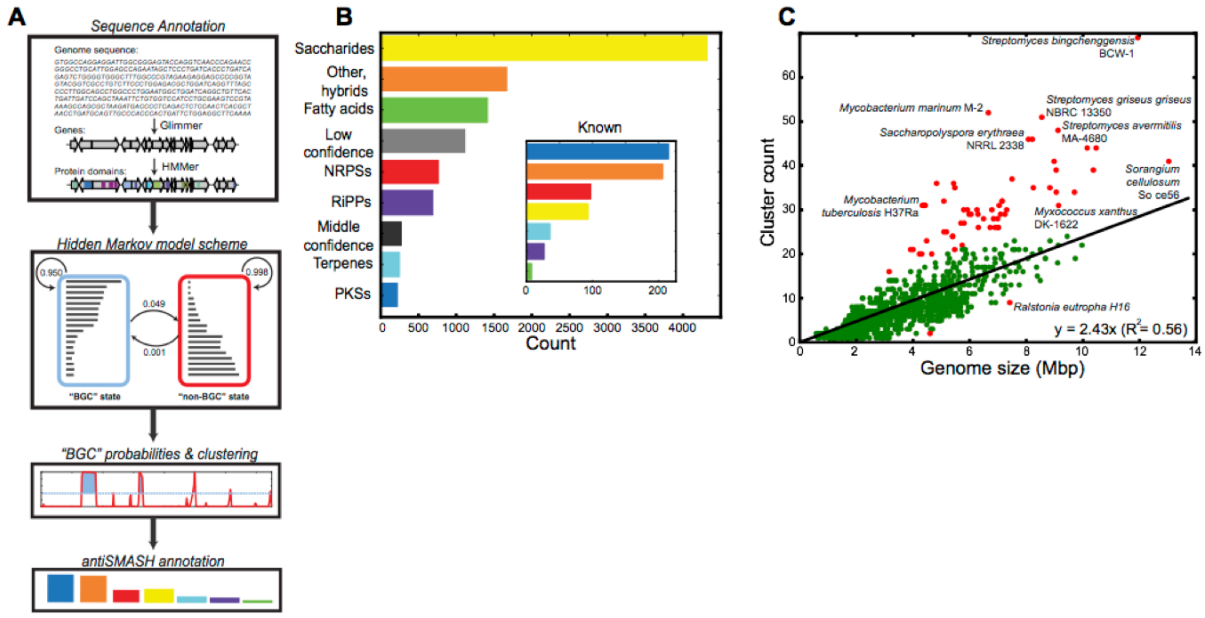

Figure 1. ClusterFinder flowchart and distribution of BGC types and counts. a, Flowchart of the four-step BGC-inference pipeline: (*i*) annotation of a genome sequence and compression to a string of consecutive Pfam domains, (*ii*) calculation of posterior probabilities of a BGC hidden state given the sequence of Pfam domains using the presented HMM scheme, (*iii*) clustering of genes that contain Pfam domain(s) with posterior probabilities of BGC hidden state above the threshold, and (*iv*) annotation of the predicted BGCs using the antiSMASH algorithm. **b,** Distribution of BGC types for known (inset) and predicted BGCs. We define low and middle confidence types as BGCs that could not be annotated to any "classic" BGC type, but do still contain Pfam domains with putative biosynthetic activity. **c,** Number of predicted BGCs by genome size. Most of the organisms follow a linear trend (the equation in the bottom-right corner); outliers (defined as having residuals >8) are colored red.

Strikingly, 46% of all predicted BGCs encode saccharides, more than twice the size of the next largest class. Some cell-associated saccharides such as lipopolysaccharides, capsular polysaccharides, and polysaccharide A are known to play key roles in microbe-host and microbe-microbe interactions, while diffusible saccharides have a range of biological activities, most notably antibacterial; however, only 13% of previously reported BGCs encode the biosynthesis of saccharides (**Supplemental Text 2**). Nearly every species harbors saccharide gene clusters, and in 59% of species, more than half of the predicted gene clusters encode saccharides. 32% of the saccharide BGCs are not closely related to any known gene cluster; these include BGCs from entirely unexplored genera (**Supplementary Figure 1**). Saccharide BGC repertoires are also surprisingly diverse: only 37% occur in the genomes of two species chosen at random from the same genus (compared to 43% for polyketides, 60% for terpenoids and 74% for fatty acids, **Supplementary Figure 2**). The abundance of novel saccharide BGC families raises the possibility that more clinically relevant saccharides such as the antidiabetic drug acarbose and the antibiotics gentamicin and avilamycin will be discovered. Another BGC class of unexpectedly large size is the one encoding ribosomally synthesized and posttranslationally modified peptides (RiPPs,(Arnison et al. 2013)). Surprisingly, RiPP BGCs are as prevalent in our data set as those encoding nonribosomal peptides (**Figure 1b**). Unexpectedly, most gene clusters (84%) belong entirely to a single class; hybrid BGCs are a small minority.

Recent case studies have shown that BGCs can be among the most polymorphic elements between closely related genomes (Jensen et al. 2007; Tobias et al. 2013), which suggests that they play key roles in ecological specialization. However, it is not clear whether the plasticity of BGC complements

is a universal phenomenon or whether it is limited to certain taxa. To investigate patterns of BGC repertoire diversity and evolution on a global scale, we analyzed BGC distributions throughout the bacterial tree of life (**Supplemental Text 3**) and used a quadratic entropy index to estimate the diversity of gene clusters among the nodes of the tree (Pavoine, Baguette, Bonsall 2010).

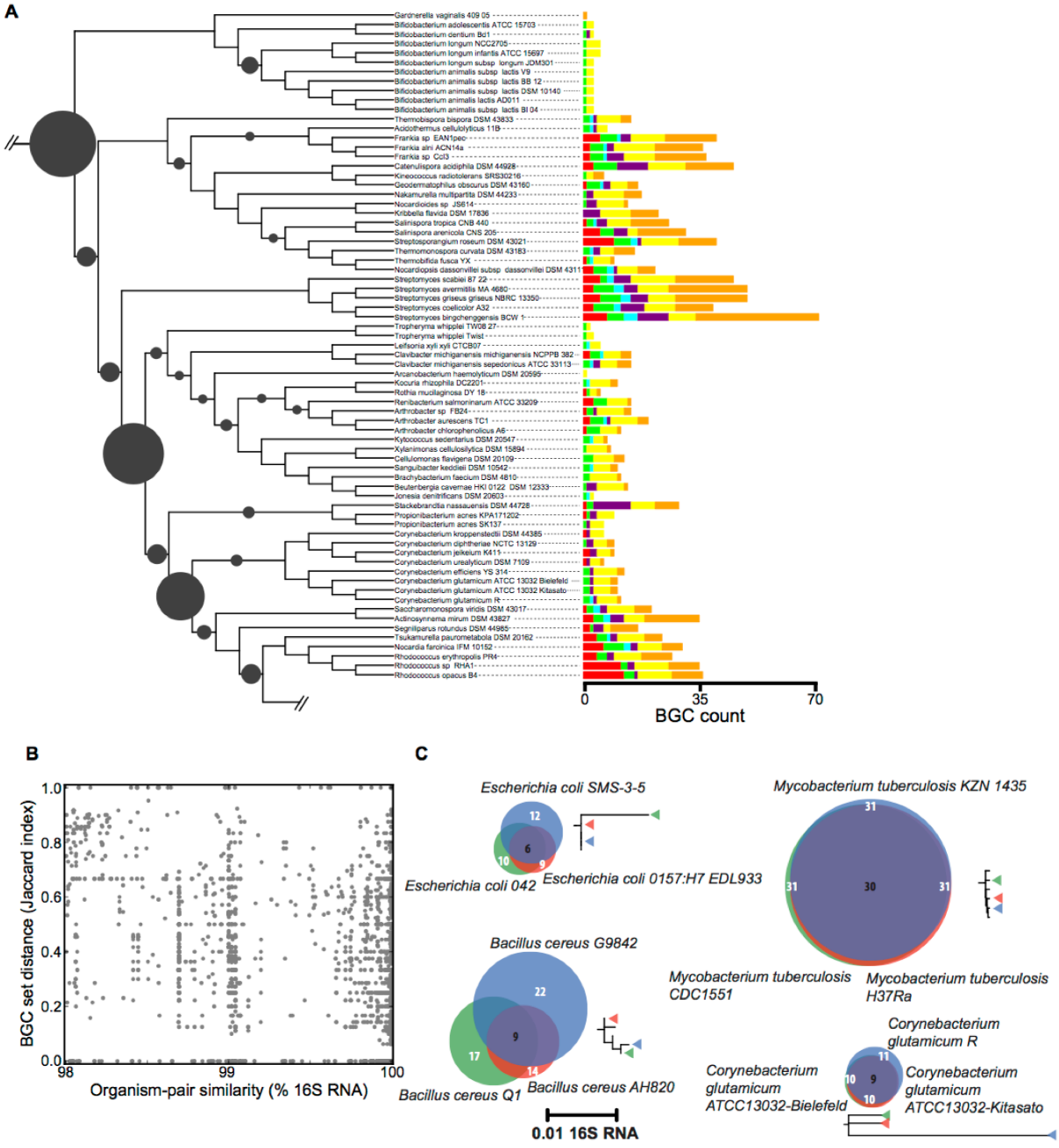

Figure 2. Diversity of BGCs is independent from phylogeny. a, Decomposition of BGC diversity among species of the phylum Actinobacteria. The diversity of each node in the phylogenetic tree is measured by the quadratic entropy index, and represented by the size of the circle (larger circle defines higher degree of diversity). Color bars at the leaf tips represent number of BGCs per species, with different colors denoting different BGC types (colors as in **Figure 1b**). Hybrid gene clusters (orange) are unusually prominent in Actinobacteria (~50%). **b**, The scatter plot shows no correlation between phylogenetic and BGC content distance for a given organism pair. **c,** The Venn diagrams show the number of BGCs shared among three different sets of closely related species. The phylogenetic tree sections to the right of the Venn diagrams are shown using the same scale.

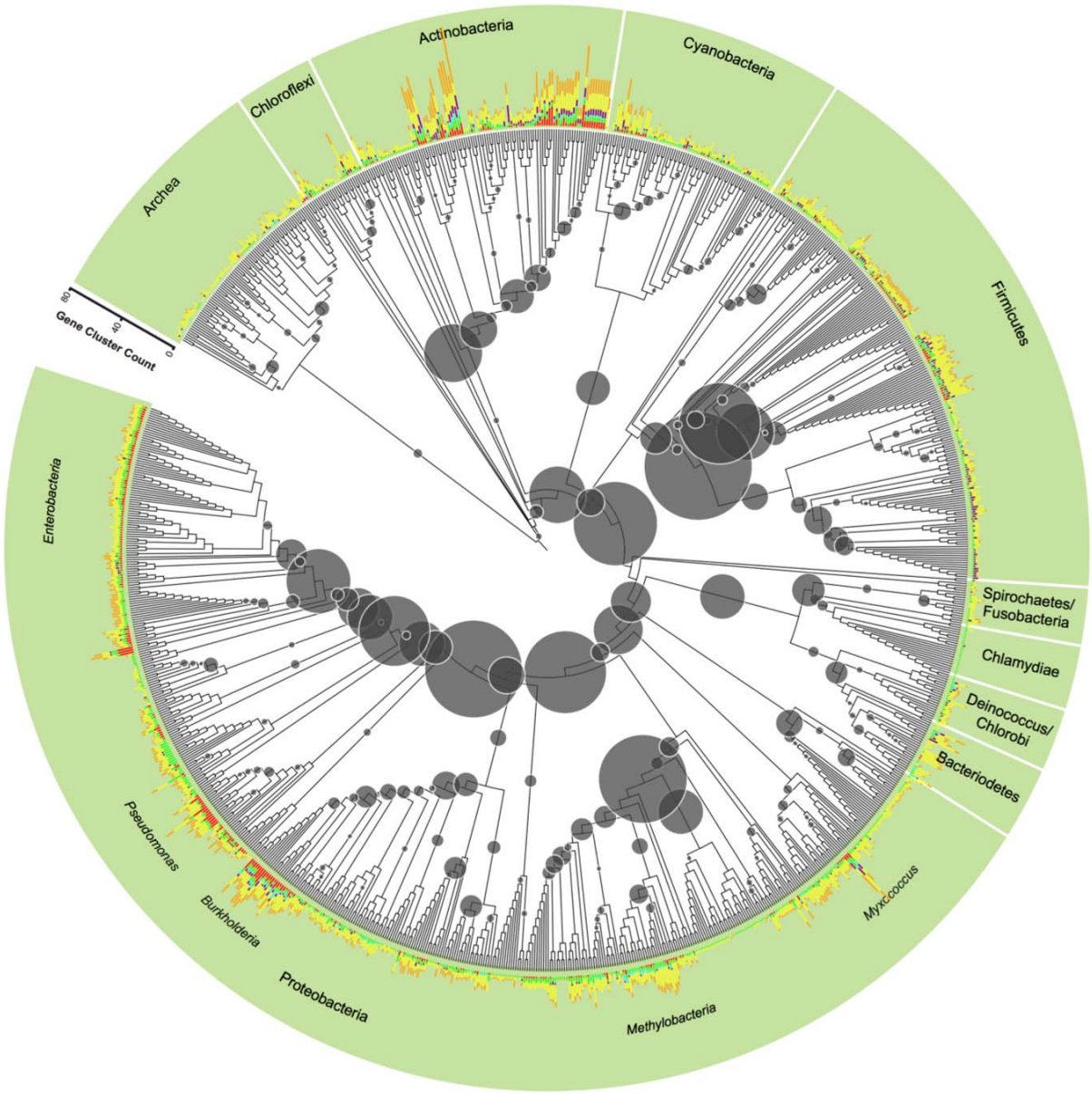

Figure 3. Decomposition of BGC diversity among all sequenced prokaryotic genomes. The diversity of each node in the phylogenetic tree is measured by the quadratic entropy index, and represented by the size of the circle (larger circle defines higher degree of diversity). Color bars at the leaf tips represent the number of BGCs per species, with different colors denoting different BGC types (colors as in **Figure 1b**).

Surprisingly, we find that the degree to which gene clusters are shared within a taxon differs markedly among bacterial taxa (**Figure 2 & 3; Supplementary Figure 3; Supplemental Text 4**). For example, while three strains of *Escherichia coli* and *Bacillus cereus* share 32% (6 out of 19) and 26% (9 out of 35) of their pan-gene-cluster complement, respectively, three strains of *Corynebacterium glutamicum* that span a comparable phylogenetic distance share 70% (9 out of 13) of their gene clusters (**Figure 2c**). Our analysis also highlights the opportunity to identify new natural products by studying underexplored taxa (Letzel, Pidot, Hertweck 2013). For example, species from the genera *Legionella* and *Coxiella* stand out as intracellular pathogens that have retained multiple BGCs in spite of their minimized genomes (**Supplementary Figure 4**), indicating a strong selective pressure for the small molecule products of these pathways.

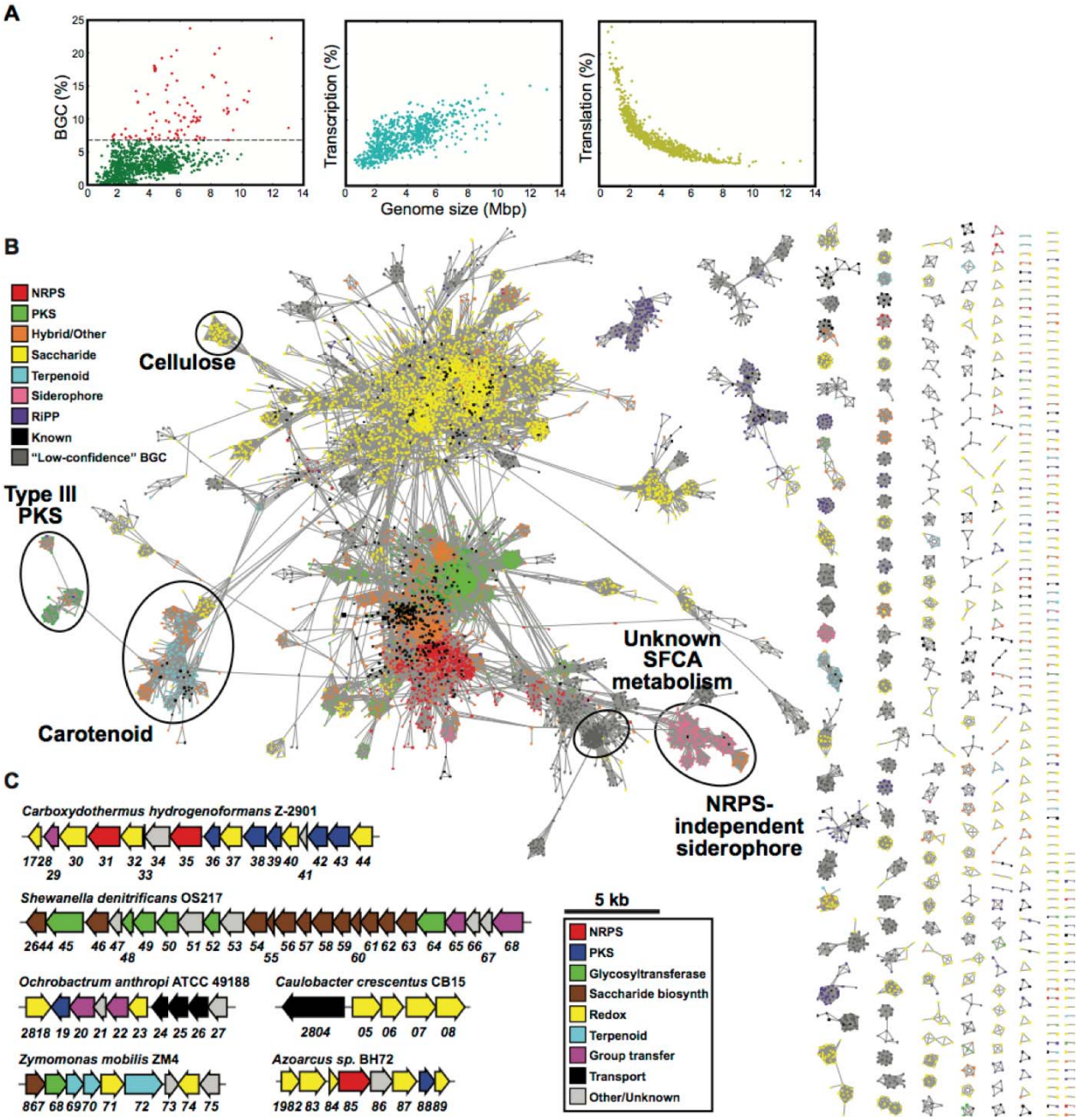

Figure 4. A systematic analysis of bacterial BGCs. a, The proportions of bacterial genomes devoted to secondary metabolite biosynthesis (left panel; 6.7% of species that devote >7.5% of their genome to biosynthesis are marked red), transcription (middle panel), and translation (right panel). **b,** Similarity network of known and putative BGCs, with the BGC similarity metric threshold at 0.5. See text for details. **c,** Examples of novel BGC families of unknown function.

The evolution of BGCs is largely independent of the phylogeny of their host genomes (Fischbach, Walsh, Clardy 2008). To systematically study relationships among BGCs, we adapted a measure of the evolutionary distance between multi-domain proteins (Lin, Zhu, Zhang 2006). We calculated an all-by-all distance matrix for the 9,421 BGCs in our high confidence set along with 1,048 manually selected BGCs in our low confidence set and the 732 members of our training set. The resulting network, the topology of which is robust to changes in the distance threshold, reveals two key findings (**Figure 4B & Supplemental Text 5**). First, one connected component harbors most of the gene clusters (72%), and is largely composed of two linked subgraphs: one dominated by saccharide

BGCs and the other a mixture of nonribosomal peptide (NRP) BGCs and polyketide/lipid BGCs, indicating that BGC from these classes share a significant number of gene families with one another. Second, there are many prominent subgraphs in which no gene clusters have been characterized; some of these BGCs may encode entirely novel chemical scaffolds. From these unexplored subgraphs, many of which include 'low-confidence' BGCs, three common themes emerge, each pointing to a putative large class of chemically novel secondary metabolites (**Figure 4c**): (*i*) There are dozens of gene cluster families ranging from 3-20 kb that harbor a 3-ketoacyl-ACP synthase (KAS) III enzyme and a diverse and varying set of auxiliary tailoring enzymes including desaturases, adenylation domains, and aminotransferases. These occur in well-studied organisms such as *Burkholderia* as well as unexplored genera such as *Anaeromyxobacter* and *Ochrobactrum*. (*ii*) There is an abundance of uncharacterized terpenoid, lipid, and glycolipid gene clusters in poorly studied genera such as *Zymomonas*, *Acetobacter*, *Nitrobacter*, and the archaeon *Sulfolobus*, which are unlike any known BGCs from these classes. (*iii*) There is a diverse set of gene clusters that are rich in redox enzymes without containing bond-forming enzymes for known compound classes, exemplified by a gene cluster consisting of four flavin-dependent halogenases and a TonB-dependent receptor from *Caulobacter.* Exploring the BGCs from these subgraphs may reveal novel chemical scaffolds, for which there is a great need in antimicrobial drug discovery (Fischbach and Walsh 2009). Experimental characterization of members of a subgraph containing >500 related BGCs has already revealed a previously unrecognized gene cluster family, widely distributed among Gram-negative bacteria, which produces a novel polyketide scaffold (Fischbach et al, in preparation).

How much secondary metabolic diversity is left to discover? Even with conservative assumptions, we estimate the total number of bacterial BGC families (such as those encoding carotenoids or calcium-dependent lipopeptides) present in the biosphere to be ~6,000 **(Supplementary Figure 5)**, less than half of which are identified in our current set of genomes (~2,400). Importantly, each of these 6,000 families will likely contain a range of molecules with distinct biological activities. Thus, for the foreseeable future, the number of gene clusters encoding molecules with distinct scaffolds will continue to rise as new genomes are sequenced.

Two questions have stoked interest in studying the evolution of BGCs: First, how does Nature modify existing gene clusters to invent new molecules? Second, do the mechanisms by which gene clusters evolve hold lessons for BGC engineering? We observed that horizontal gene transfer, insertions/deletions, rearrangements and duplications occur at unusually high rates in BGCs (**Supplemental Text 6**). Intriguingly, phylogenetic profiling showed that many common sets of domains within BGCs – here termed sub-clusters – appear to evolve in a correlated fashion: 884 different motifs (out of 7,641 found) were shown to co-evolve significantly more often than not ($P < 0.001$), based on the χ^2 test. These motifs comprise 591 different Pfam domains and have an average length of 5.3 domains (**Supplementary Table I**).

To further explore the role of sub-clusters in the evolution of BGCs, we compiled a set of 34 BGCs that are rich in sub-clusters for which links to a specific chemical moiety have already been established in the literature. We constructed a network in which the nodes represent BGCs and each edge denotes a sub-cluster that a pair of BGCs has in common (**Figure 5**; see also **Supplemental Text 6**). Three observations were particularly notable (**Figure 5**). First, >60% of the coding capacity of some BGCs (e.g., those encoding vancomycin and rubradirin) is composed of sub-clusters. This supports a "bricks and mortar" model of gene cluster evolution in which gene clusters are composed

of large, modular “bricks” (sub-clusters) that encode key building blocks and individual genes (the “mortar”) that encode functions such as tailoring, regulation and transport.

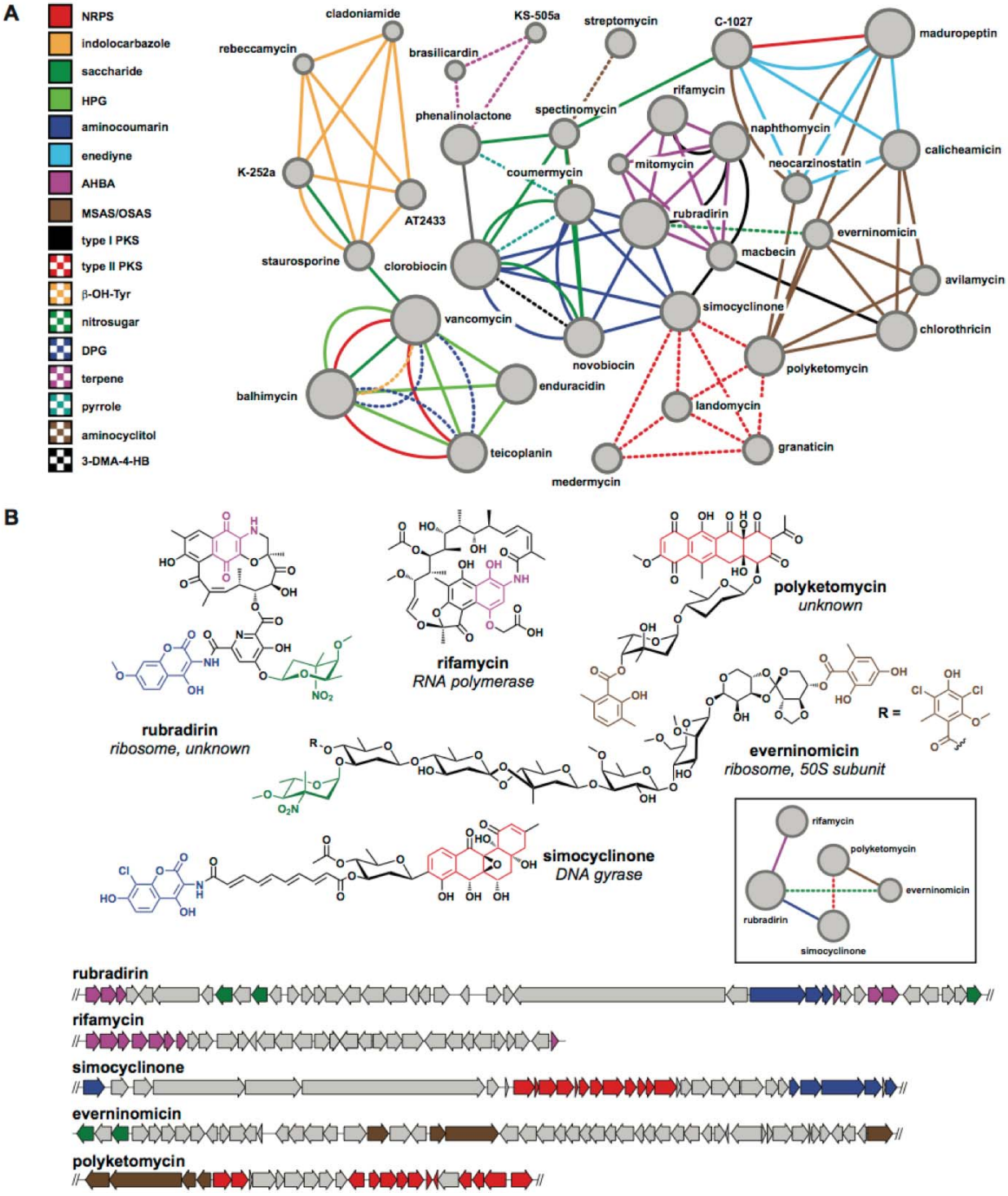

Figure 5. The successive merger of smaller sub-clusters plays a central role in the evolution of BGCs. a, Network of sub-clusters shared among 34 known BGCs. Nodes represent BGCs, and node size indicates the number of sub-clusters present in the gene cluster that are shared with other BGCs within the network. Edges represent shared sub-clusters, coded by color. **b,** A sub-network from **a** showing the shared sub-clusters among the BGCs for rubradirin, rifamycin, simocyclinone, everninomicin, and polyketomycin, as well as the chemical moieties encoded by the sub-clusters.

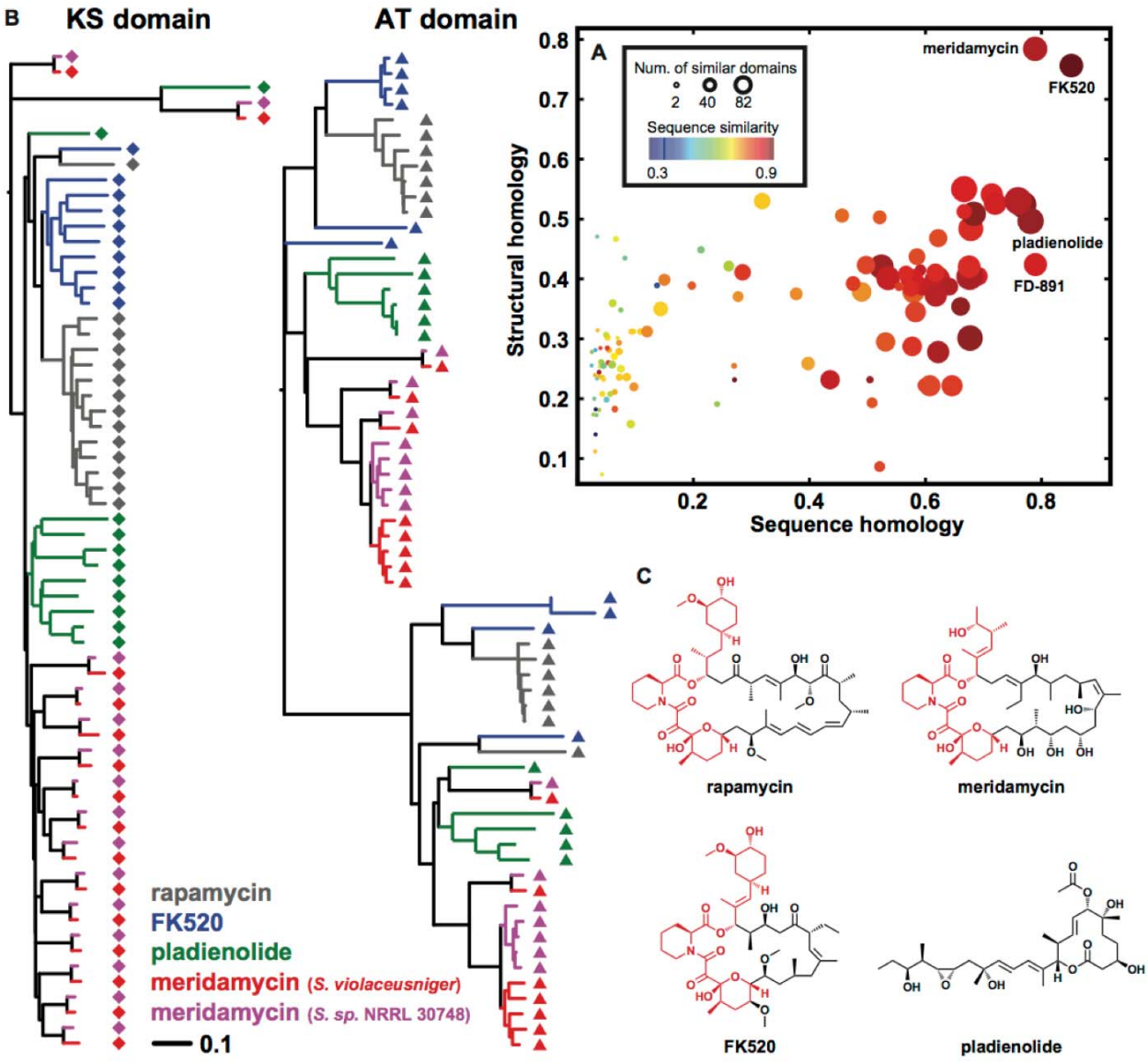

Figure 6. Unexpected evolutionary relationships within the rapamycin family. a, Distinct scaffolds from related BGCs. The scatter plot shows the relationship between the sequence homology of a pair of BGCs (x-axis) and the structural homology of their small molecule products (y-axis). Each circle represents a gene cluster and its small molecule product, and this plot shows their relationship to rapamycin and its BGC (scatter plots for other reference BGCs are shown in the Supplementary Information). Meridamycin and FK520 are closely related to rapamycin, as are their BGCs. While the pladienolide BGC is closely related to the rapamycin BGC, pladienolide itself is not closely related – it has a distinct scaffold and protein target. Structural homology is estimated by the Tanimoto coefficient, while sequence homology is represented as the Jaccard index defined on pairs of Pfam domains that share sequence identities within the top 10th percentile of all-pair sequence identities. The number of domain pairs that share sequence identities within the top 10th percentile and sequence identity of all domain pairs are shown as point sizes and colors, respectively. **b,** The role of concerted evolution in homogenizing domains within a BGC. Phylogenetic trees of KS and AT domains from the rapamycin, FK520, meridamycin, and pladieonolide BGCs are shown. The KS and AT sequences cluster into BGC-specific clades, even for the AT domains of two different clusters encoding the same compound (meridamycin), showing the ability of concerted evolution to homogenize domains within a BGC. **c,** Chemical structures of rapamycin, meridamycin, FK520 and pladienolide. The sub-structure shared among rapamycin, meridamycin and FK520 is colored red.

Second, the same sub-cluster commonly appears in otherwise unrelated BGCs, and multiple unrelated sub-clusters can be found in a single parent gene cluster, indicating that sub-clusters are independent evolutionary entities. Third, sub-clusters are not static; they are loosely organized

around a core set of genes, but gene gain/loss leads to chemical changes in the corresponding part structure: for example, gene clusters encoding molecules such as everninomicin, simocyclinone, and polyketomycin have different variants of deoxysugar sub-clusters, which leads to subtle variations in the final chemical structures.

To analyze how gene clusters encoding large assembly-line biosynthesis machineries may evolve from producing one scaffold to another, we calculated the proportion and similarity of Pfam domains shared between a pair of BGCs using multiple sequence alignments for each Pfam domain **(Figure 6)**. The results reveal unexpected evolutionary connections among natural products (see also **Supplemental Text 7**). For example, the *Streptomyces* gene cluster encoding the lipopeptide antibiotic daptomycin is surprisingly similar to *Mycobacterium* glycopeptidolipid (GPL) gene clusters (**Supplementary Figure 6**). Likewise, one of the strongest matches for the gene cluster encoding the immunosuppressant rapamycin, apart from the closely related FK520 and meridamycin BGCs, was the gene cluster for pladienolide, a polyketide of unrelated structure with a distinct biological activity (inhibition of the splicing factor SF3b instead of TOR). Strikingly, based on phylogenetic trees of their constituent ketosynthase (KS) and acyltransferase (AT) domains, the meridamycin gene cluster is more closely related to the pladienolide BGC than to those encoding rapamycin and FK520, the molecules to which it is often compared (**Figure 6**). These examples demonstrate how closely related domains can be reconfigured by recombination to yield a new scaffold that is chemically and biologically distinct.

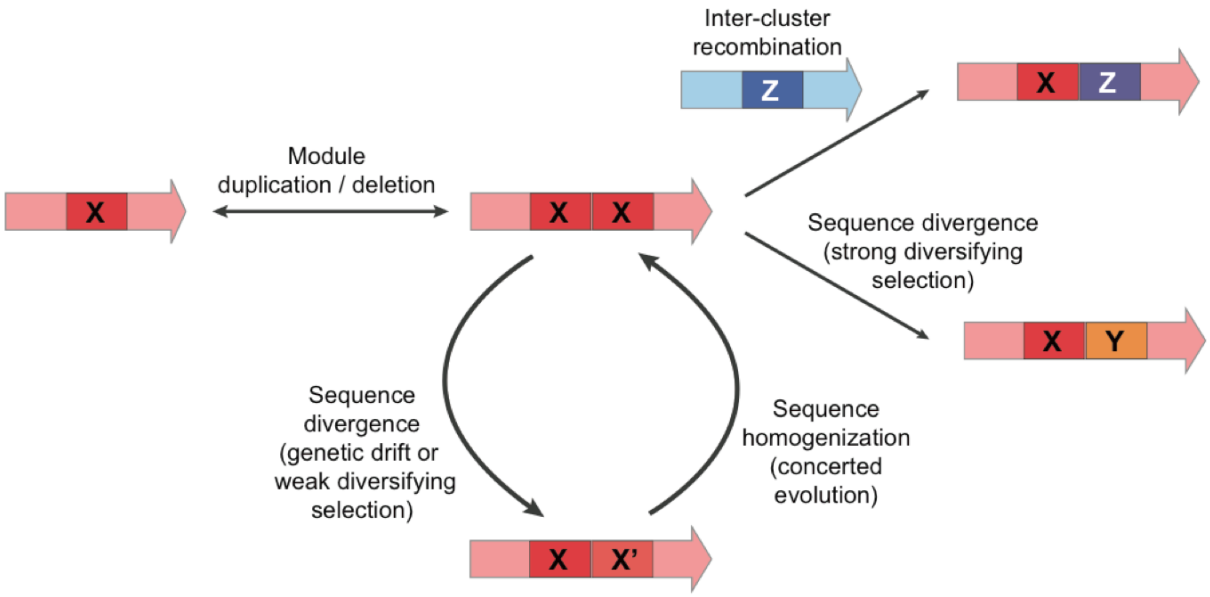

Figure 7. Qualitative model for the evolution of NRPS/PKS domains. After modules are duplicated, they may get 'trapped' in a cycle in which small sequence divergences are counterbalanced by internal recombinations that drive concerted evolution. Through strong diversifying selection (or sufficient drift), domains may break out of this cycle towards domain sequences that are protected from concerted evolution by functional divergence and subsequent stabilizing selection on the new function, or by reduced internal recombination rates due to larger sequence differences between the domains. Sequences may diverge either by mutation or recombination with other gene clusters (or with other modules in the same gene cluster).

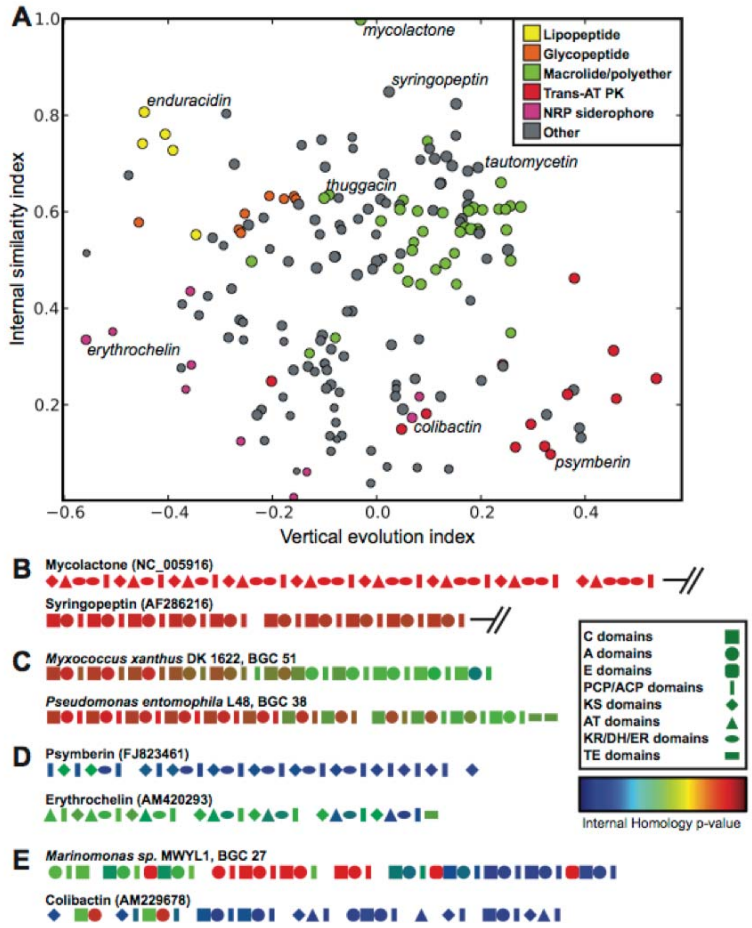

Figure 8. Four distinct modes of evolution for PKS and NRPS BGCs. a, Scatter plot showing two features of BGCs – internal similarity index and vertical evolution index – that, of the 25 measured, describe the most variation from a principal component analysis. Colors indicate distinct classes of BGCs. **b-e,** Domain architecture plots of PKSs and NRPSs show distinct modes of evolution: **b,** Internal duplication with concerted evolution; **c,** N-terminal additions by module duplication and recombination; **d,** domain swapping with other BGCs; and **e,** mixed evolution. Geometric shapes indicate domain types (see legend); domain colors indicate the internal homology p-value of each domain to its nearest neighbor within the same gene cluster.

The phylogenetic trees of KS and AT domains revealed another unexpected finding: in spite of the structural similarity of rapamycin and FK520, 63% of the constituent domains of their polyketide synthases (PKSs) cluster into entirely separate clades. This pattern of homology can be explained by 'concerted evolution', the homogenization of DNA sequences within a given repetitive family caused by high rates of internal recombination (Liao 1999; Santoyo and Romero 2005). Previous phylogenetic analyses of PKS domains have observed BGC-specific clades of PKS domains (Jenke-Kodama, Borner, Dittmann 2006; Zucko et al. 2012), but not to the extent observed here for such closely related gene clusters. The fact that such a strong pattern is even observed for the AT domains of two different gene clusters that encode the same molecule, meridamycin, shows that the underlying process may operate on much shorter evolutionary timescales than previously thought, and that recombination can remove almost all traces of vertical evolution of these PKS modules. Concerted evolution is not peculiar to the rapamycin family; we observed it to different extents in closely related macrolide (KS and AT domains) and ansamycin and polyene (AT domains) gene clusters, but—corroborating earlier observations (Nguyen et al. 2008)—not in trans-AT PKSs

(**Supplementary Figure 7**). Certain nonribosomal peptide synthetase (NRPS) gene clusters (such as lipopeptide BGCs, but not glycopeptide BGCs) also show signs of concerted evolution (**Supplementary Figure 7**). Our qualitative model of PKS/NRPS evolution (**Figure 7**), which summarizes the interplay of concerted evolution with other evolutionary mechanisms, is relevant to PKS/NRPS engineering efforts: the highly homologous sets of domains generated by concerted evolution are more likely to be mutually interoperable than domain sets chosen at random, and might therefore be of utility in future engineering efforts.

To understand more generally how PKS and NRPS BGCs evolve, we set out to measure the contributions of concerted evolution, duplication, and divergence to their evolution. We first collected and quantified 25 different features describing the nature of gene cluster sequences and the relationships among their constituent domains. Principal component analysis (PCA) showed that these features describe much of the variation that distinguishes many of the well-known gene cluster families (**Supplementary Figures 8 & 9**). Two features in particular, the 'internal similarity index' and the 'vertical evolution index', explain much of the variation in terms of the modes of evolution of different classes of gene clusters (**Figure 8a**). Zooming in to the level of individual domains (**Supplementary Figure 10**), we find that there are four primary mechanisms by which NRPS and PKS BGCs evolve (**Figure 8b-e**): duplication and concerted evolution, N-terminal acquisition of new modules by duplication or domain swapping, domain swapping with other BGCs, and mixed modes of evolution. This finding suggests that it would be advantageous to engineer BGCs in a subclass-specific manner, introducing changes at those positions in the enzymatic assembly line that have proven flexible during their evolution.

Our analysis lays the groundwork (**Supplemental Text 8**) for experimentally mining underexplored taxa, identifying unknown classes of BGCs encoding novel chemical scaffolds, and developing new approaches to BGC engineering informed by the mechanisms by which BGCs evolve naturally (Medema et al. 2011a).

Methods Summary

A set of 1154 complete genome sequences was obtained from JGI-IMG (Markowitz et al. 2012), version 3.2 (08/17/2010). The ClusterFinder prediction algorithm for BGC identification is a two-state Hidden Markov Model (HMM), with one hidden state corresponding to biosynthetic gene clusters (BGC state) and a second hidden state corresponding to the rest of the genome (non-BCG state). The training set consisted of 677 experimentally characterized gene clusters for the BGC state and non-BGC regions from 100 randomly selected genomes for the non-BGC state, defined as those regions without significant sequence similarity to the training set sequences (Pfam domain similarities with E-value > 1e-10). The algorithm was validated by comparison to 10 manually annotated bacterial genomes, and by assessing performance on 74 experimentally characterized BGCs outside the training set. Annotation of BGCs was performed using profile HMMs (pHMMs) from antiSMASH (Medema et al. 2011b), supplemented by manually designed libraries of pHMMs for fatty acid and saccharide BGCs (see SI Methods).

Taxonomic classifications of organisms were obtained from NCBI Taxonomy (Federhen 2012). Species phylogenies were constructed using the neighbor-joining method on 16S rRNA marker

sequences from the corresponding genomes from JGI-IMG. Estimates of within-taxon variation across the tree were calculated using the quadratic entropy index (Pavoine, Baguette, Bonsall 2010). Protein sequence phylogenies were performed using the neighbor-joining method, with 100 bootstrap replicates.

BGC similarity networks were calculated using a modified version of the distance metric from Lin and coworkers (Lin, Zhu, Zhang 2006) for multi-domain proteins. BGC families were calculated with a Lin similarity threshold of 0.5 and MCL clustering with I = 2.0. Sub-cluster similarity was measured using a blastp comparison with type-specific thresholds (see SI Methods). Structural similarities among small molecules were estimated with the Tanimoto coefficient.

Full Methods and associated references are available in the Supplementary Information.

Acknowledgements

This work was supported by an HHMI Predoctoral Fellowship (PC), a Boehringer Ingelheim Fonds travel grant (MHM), Grant 10463 from the GenBiotics programme of the Dutch Technology Foundation STW to ET (MHM), an NWO-Vidi fellowship (RB), a Medical Research Program Grant from the W.M. Keck Foundation (MAF), a Fellowship for Science and Engineering from the David and Lucile Packard Foundation (MAF), DARPA award HR0011-12-C-0067 (MAF), and NIH grants OD007290, AI101018, AI101722 and GM081879 (MAF). This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases National Institutes of Health, Department of Health and Human Services, under Contract No.: HHSN272200900018C.

Author Contributions

P.C., M.H.M. and M.A.F. designed the research, analyzed the data and wrote the paper, with substantial input from R.B., E.T., J.C., and A.S. P.C. and M.H.M. performed the research. K.M. and A.P. provided input data and data integration into the JGI-IMG database. L.C.W.B., P.A.G., M.K., B.W.B., and M.A.F. designed an earlier version of the gene cluster identification algorithm that served as a model for the current version.

Supplementary Material

Supplementary figures and tables can be downloaded from http://rdmy.info/ch10

Supplemental Text

1. Systematic identification of gene clusters from bacterial genomes

A total of 1154 complete bacterial genomes were analyzed. Draft genomes were not included in the analysis, because BGCs are often highly fragmented in their assemblies. Gaps in draft assemblies occur predominantly at genes encoding large biosynthetic enzymes (Klassen and Currie 2012).

Our detection method, ClusterFinder, is based on a training set of 732 BGCs with known small molecule products that we compiled and curated manually (**Supplementary Table II**). Its identification algorithm is a Hidden Markov Model (HMM) that processes a string of contiguous Pfam domains that corresponds to a genome sequence and assigns each domain a probability of being a part of a biosynthetic gene cluster (BGC) (**Figure 1a;** see **SI Methods** for an explanation of how we validated ClusterFinder). We chose not to train separate HMMs for specific gene cluster classes (e.g., polyketides or terpenes), since these narrower HMMs would be less effective at identifying hybrid and novel classes of BGCs. Since ClusterFinder is based solely on Pfam domain frequencies, and novel arrangements of the same enzyme families can yield new BGC classes (Trefzer et al. 2002), ClusterFinder exhibits relatively little training set bias and is capable of identifying new classes of gene clusters, as discussed below.

To filter and analyze predicted gene clusters, we merged ClusterFinder's results with those of a second gene cluster identification algorithm, antiSMASH, which identifies and annotates gene clusters based on a complementary strategy: a hierarchical logic of conserved protein domains that are characteristic of one of ~20 gene cluster classes (Medema et al. 2011b). Our probability threshold of 0.4 was chosen to keep the false-positive ratio below 5% (an estimate based on a comparison of ClusterFinder results and manual annotations of 10 genomes). The true-positive ratio at this threshold is 55%.

2. Saccharides are the largest class of gene clusters

We began our analysis by grouping the 9,421 high-confidence gene clusters into classes based on the presence of characteristic protein domains and asking how many of each class we recovered (**Figure 1b**). The prevalence of certain biosynthetic classes in the entire dataset could be compared with their prevalence in experimentally characterized gene clusters using our training set. This set of 732 experimentally characterized gene clusters is nearly exhaustive and was compiled in an unbiased manner, so it is a reasonable proxy for measuring how well each gene cluster class has been studied.

The predominance of saccharide gene clusters illuminates families of molecules that are not typically thought of as natural products. For example, 23% of the saccharide gene clusters are predicted to encode lipopolysaccharides and 3% capsular polysaccharides. These cell wall-mounted molecules play important roles in host-microbe and microbe-microbe interactions, and small changes in their structure can lead to large changes in their function. Other saccharides have antibacterial activity. A recently discovered saccharide BGC (with an average ClusterFinder probability of 0.93) has been found to encode saccharomicin, a member of a novel family of heptadecaglycoside antibiotics with potent activity against Gram-positive pathogens (Strobel et al. 2012).

Besides saccharide gene clusters, several other gene cluster types are notable as well. Gene clusters encoding ribosomally synthesized and posttranslationally modified peptide natural products (recently termed RiPPs (Arnison et al. 2013)) are found in much larger numbers than polyketides and terpenes. RiPPs are difficult to detect because of their immense architectural diversity (Arnison et al. 2013); as a result, they are the most likely class to be underestimated by our approach. Consequently, gene clusters for RiPPs may be among the most widely distributed categories in bacterial genomes. Finally, we also detected and manually curated around 1,500 gene clusters (subdivided into low and middle confidence categories, see Methods) that have all the hallmarks of being BGCs, but do not clearly fall into any known class of BGCs. These provide a promising set of candidate BGCs that may lead to the discovery of novel chemical scaffolds, for which there is great need in current drug development approaches (Fischbach and Walsh 2009).

3. Prolific producers harbor exceptionally large complements of gene clusters

We addressed the question of how a bacterium's genome size relates to its biosynthetic capacity. Similar to a result from an earlier report (Donadio, Monciardini, Sosio 2007), we find that bacteria have an average of 2.4 gene clusters per Mb (SE = 0.03 and 0.10, simple least squares linear regression and generalized least squares linear regression corrected for phylogeny, respectively) (**Figure 1c**). Strikingly, however, certain strains are clear outliers in that they have more than the average number of gene clusters per Mb (defined as having residuals >8, 5.0% of the total). The scarcity of low-end outliers suggests that nearly all bacterial species harbor at least a minimal complement of biosynthetic gene clusters.

Likewise, we find that while the average species devotes 3.7% ± 3.1% of its genome to BGCs, a largely overlapping group of outlier species devote >7.5% of their genomes to natural product biosynthesis (defined as >1 SD above the mean, 6.7% of the total). This is comparable to the mean fraction of a bacterial genome devoted to transcription (7.2%) and translation (8.5%) (**Figure 4a**). One outlier, *Streptomyces bingchenggensis*, devotes a remarkable 22% of its genome to secondary metabolites; in aggregate, this strain's gene clusters (2.65 Mb) are larger than the entire genome of every sequenced strain of *Streptococcus*. The aggregate gene clusters of a less extreme strain, *Streptomyces griseus* (1.77 Mb), still dwarf most *Helicobacter* genomes.

Many of the outliers are strains of *Streptomyces*, *Myxococcus*, *Sorangium*, and *Burkholderia*. Our results suggest that it is probably no coincidence that these genera have long been known for their prolific production of natural products, since they harbor an exceptionally large complement of gene clusters. Importantly, other outliers are from genera that, to our knowledge, have not yet been mined for natural products: *Gloeobacter*, *Methylobacterium*, *Shewanella*, and *Teredinibacter*. In general, there is a vast discrepancy in phylogenetic distribution between experimentally characterized gene clusters in our training set and our set of predicted gene clusters (**Supplementary Figure 11**). Further highlighting the opportunity to identify new molecules by studying underexplored taxa, species from the genera *Legionella* and *Coxiella* stand out as intracellular pathogens that have retained multiple BGCs in spite of their reductive genome evolution (**Supplementary Figure 4**), indicating a strong selective pressure for the small molecule products of these gene clusters.

We next mined metadata on BGC-harboring organisms from the NCBI BioProject/BioSample databases (Barrett et al. 2012) to identify correlations between the numbers of gene clusters in a genome and the ecology or lifestyle of a microbe. We find that organisms that display a large degree of multicellularity, occur in terrestrial habitats, form endospores and/or have an aerobic lifestyle have more gene clusters on average than organisms that do not exhibit these features (**Supplementary Table III**). Nonetheless, the biosynthetic potential from species without these features should not be underestimated: even though anaerobes have on average six times fewer gene clusters, these taxa have not been well explored and therefore hold great promise for further study (Letzel, Pidot, Hertweck 2013).

In a more general sense, we observe that the length of a bacterial genome correlates best with the size of coding regions for transcription-associated genes as well as primary and secondary metabolism, while the size of the coding regions for other functional categories remains constant (**Figure 4a** and **Supplementary Figure 12**). Thus, it would appear that bacterial genomes expand largely to increase their gene complements for transcription, primary metabolism, and secondary metabolism.

4. The relationship between phylogeny and gene clusters varies tremendously across the bacterial tree of life

The phylogenetic distribution of BGCs is a key factor in understanding their biological roles. If related species harbor similar BGCs, then their small molecule products could underlie phenotypes common to the taxon. Alternatively, if related species harbor different gene clusters, then these elements could play an important role in ecological specialization. Evidence for the latter has come from recent reports showing that genomes of *Mycobacterium* and *Bacillus* are 92-98% similar at the nucleotide level, yet differ markedly in their complement of gene clusters (Rückert et al. 2011; Tobias et al. 2013). However, it is not clear whether this phenomenon is general or specific to these taxa.

To answer this question, we used a quadratic entropy index to illustrate how the diversity of gene clusters can be decomposed among the nodes of the phylogenetic tree (Pavoine, Baguette, Bonsall 2010). This methodology allowed us to determine gene cluster diversity at internal nodes at different depths in the phylogeny (**Figure 2a & 3**). Surprisingly, we find that the degree to which gene clusters are shared within a taxon differs markedly among bacterial taxa. For example, while species of Actinobacteria and Deltaproteobacteria harbor widely varying gene cluster complements, species of Enterobacteria and Firmicutes harbor similar numbers and types of gene clusters. But even BGC repertoires of closely related strains from the latter (sub)phyla can display notable differences: for instance, *Bacillus subtilis* ATCC 6633 (Zeigler 2011) shares the bacillibactin, bacillaene, surfactin, subtilosin and bacilysin gene clusters with the common laboratory strain *B. subtilis* 168. However, *B. subtilis* ATCC 6633 harbors a mycosubtilin gene cluster in place of the plipastatin gene cluster found in *B. subtilis* 168 -- two nonribosomal peptide gene clusters of similar size that produce small molecule products in distinct families (**Figure 2b**). In addition, *B. subtilis* ATCC 6633 harbors the gene clusters for subtilin and rhizocticin, while *B. subtilis* 168 encodes the cannibalistic SDP and SKF peptides (Liu et al. 2010).

In general, we find that the diversity of BGCs does not appear to be skewed towards the root or the leaves of the phylogenetic tree (**Supplementary Figure 3**), indicating an ongoing process of gene cluster diversification. We observe many nodes of high diversity in the tree closer to the leaves, pointing to evolution independent of phylogeny, perhaps indicative of ecologically driven diversification.

5. A global map of biosynthesis based on a gene cluster distance metric

In order to draw a global network that shows the mutual evolutionary relationships between all the BGCs in our dataset, we used the distance metric of Lin et al. (2006). The distance metric has two components: the first is based on the Jaccard coefficient and measures the similarity between the gene families included in each gene cluster, and the second represents the copy number variation of gene families between the two clusters. We validated that the distance metric works in this setting by using it to measure the distances among every pair of gene clusters in our training set; we confirmed that the gene clusters for a natural product family (e.g., glycopeptides and lipopeptides) are collectively more similar to each other than to other related clusters (e.g., other nonribosomal peptides) (**Supplementary Figure 13**). In addition, we created a high-resolution variant of the distance metric in which Pfam domain sequence similarity was also taken into account (see **Methods**). Since this version of the algorithm is more computationally intensive, we only applied it to the network of known BGCs.

We constructed and manually inspected the networks that result from making our threshold more or less stringent. The network structure shown in **Figure 4b** is robust to small variations in the clustering threshold (+/- 0.1). Larger variations yielded networks that were almost fully connected or highly dissociated, neither of which provide biological insight into the large-scale relationships among gene cluster classes. While the network in **Figure 4b** may appear densely connected, it contains just 0.6% of all possible edges (388,411 out of 63,286,875).

In the network displayed in **Figure 4b**, saccharides, nonribosomal peptides and polyketides/lipids feature prominently. Other prominent BGC clusters include terpenes, NRP-independent siderophores, and a set of unknown gene clusters that are connected to the NRPS/PKS BGC cluster (see below). The BGC clusters are densely connected by edges, indicating hybridism among gene cluster families; not surprisingly, the most prevalent hybrids are NRPS-PKS and PKS-saccharide (**Figure 4b**).

Unexpectedly, however, most gene clusters (84%) belong entirely to a single class. Hybrids therefore comprise a much larger proportion of known gene clusters than predicted gene clusters, suggesting that they may have been oversampled by experimental efforts to date. The distribution of known gene clusters in the network (black dots) is non-uniform, suggesting that efforts to experimentally characterize gene clusters have been biased toward specific BGC classes.

Since the gene clusters for RiPPs do not share core domains (Arnison et al. 2013), their biosynthetic loci do not cluster in the network; rather, they constitute distinct clusters for different RiPP subclasses (e.g., lantipeptides, thiopeptides). This corroborates their mode of evolution: RiPP BGCs

tend to be smaller and more diverse, and commonly incorporate tailoring genes from the other gene cluster classes.

Interestingly, the topology of the network offers important insights into BGC evolution. The BGC similarity graph is a small-world, scale-free network (Barabasi and Oltvai 2004): the exponent of the degree distribution, the average shortest path, and the average clustering coefficient are 1.66 ± 0.07, 1.11, and 0.69, respectively (**Supplementary Figure 14**). In small-world networks, the path between two nodes selected at random is unusually short on average; this means that for most pairs of unrelated BGCs, there will be a third gene cluster that shares a substantial number of genes with each of them. The unusually gradual descent of a node degree distribution indicates that if a new node is added to the graph, an unusually large number of edges is likely to be added (Seyed-Allaei, Bianconi, Marsili 2006). Both of these characteristics are consistent with the view that the total set of BGCs is composed of a finite set of parts used in many different arrangements and contexts. Interestingly, highly linked nodes are unusually abundant (429 hubs with more than 200 links). Some of these nodes are small BGCs that are similar to common sub-clusters from larger BGCs, suggesting that such larger BGCs often evolve through the merger of smaller BGC modules.

6. Functionally distinct sub-clusters are the basic evolutionary building blocks of BGCs

To begin addressing the questions how Nature modifies BGCs during evolution and whether these evolutionary patterns hold lessons for BGC engineering, we first set out to quantify the extent to which a key phenomenon in bacterial genome diversification – horizontal gene transfer (HGT) – has played a role in BGC evolution. While several focused case studies (Nguyen et al. 2008) have pointed to a key role for HGT in the evolution of secondary metabolism, no effort has been made to study its effects across gene clusters systematically. Our results show that HGT is a global phenomenon that has a pervasive effect on the evolution of gene clusters and gene cluster repertoires: we observed that the closest homologs of BGC genes often originate from much more distantly related organisms than the closest homologs of genes encoding the translation apparatus or genes from the tryptophan and histidine biosynthesis operons (**Figure 9a**). Moreover, we found that BGCs are regularly transferred between taxonomic orders: while only 3% (18 cases) of 16S rRNA sequence pairs that share regions of >1000 bp with >70% sequence identity were found between species of different taxonomic orders that are more than 0.2 phylogenetic distance units apart, we found the same was true for 60% (719 cases) of BGC sequence pairs (**Supplementary Figure 15**).

In light of the prominent role of HGT in secondary metabolism, we decided to examine gene cluster evolution at higher resolution. To identify and quantify recent evolutionary events in BGC evolution, we compared gene clusters that share regions of >1000 bp with >70% identity (**Supplementary Table IV**). To provide a basis for comparison, we calculated the same evolutionary events for pairs of gene clusters encoding the translation apparatus, correcting for the smaller size of the latter gene clusters (5-15 kb) by imposing a size limit of 15 kb on the BGCs to be compared. From the 10,096 BGC pairs meeting these criteria, 1,750 had a rearrangement, 1,140 had an indel, and 135 had a duplication, each of which were far more common than the corresponding evolutionary events in gene clusters encoding the translation apparatus (**Figure 9b**). Interestingly, while indels and rearrangements could be detected in ~16% and ~19% of gene clusters of all sizes, duplications are

found far more commonly in gene clusters >40 kb (7.6%) than in gene clusters 10-20 kb (0.3%), suggesting a possible role for duplication and divergence in the evolution of large gene clusters.

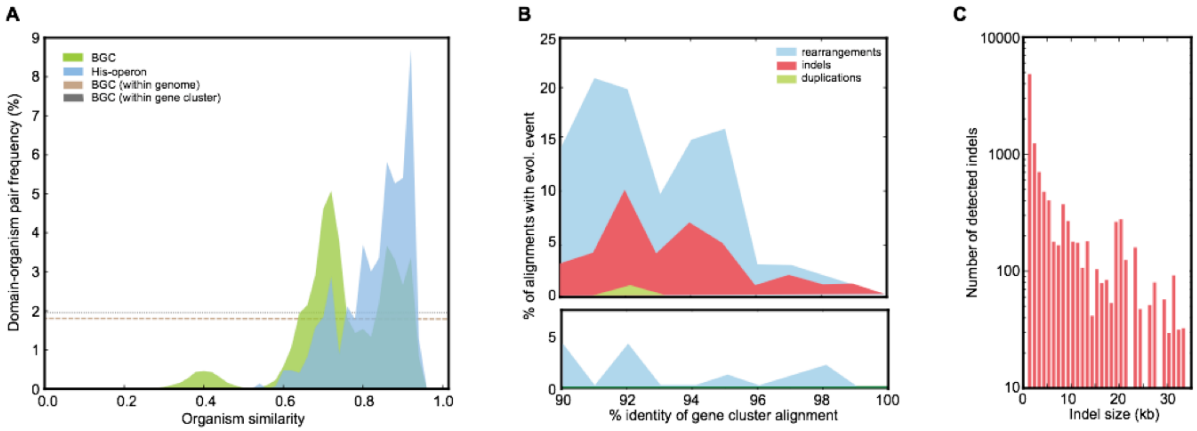

Figure 9. The evolution of BGCs differs from the evolution of ribosomal genes and primary metabolism. a, Distributions of the best matching sequence homologs with respect to organism similarity (based on 16S rRNA) for predicted BGCs and histidine operons suggest significant differences in the ways they evolve. **b**, Number of detected rearrangements, indels and duplications plotted against the average percent identity in the aligned gene cluster pairs from which the events were deduced for predicted BGCs (top) and ribosomal gene clusters (bottom). **c**, Size distribution of inserted/deleted fragments during recent gene cluster evolution, based on the indel analysis.

While the percentage of gene cluster pairs related by an indel is independent of gene cluster size, the sizes of the indels varied with a power law distribution, with a large tail that includes 195 indels of 10 kb or more (**Figure 9c**). As expected, these large indels are more commonly found in larger gene clusters, where they indicate the merger of one sub-cluster with another (see examples in **Supplementary Figure 16**).

As these results corroborate our observations from the network analysis suggesting the sharing of 'sub-clusters' between many BGCs, we decided to look more closely at the composition and evolution of sub-clusters, by detecting sub-clusters automatically in all BGCs and semi-manually in known BGCs. The automatically detected sub-clusters (using phylogenetic profiling; see **Methods**) included many well-known and widely conserved motifs that appear to be linked to specific sub-functionalities of gene clusters, such as precursor biosynthesis, transport or synthesis of a specific chemical moiety, and motifs belonging to modular BGC architectures of NRPSs and PKSs (e.g., C-A-T and KS-AT-T).

For our analysis of sub-clusters in known BGCs, we required as a cut-off for sub-cluster sharing that at least 75% of the genes of a given sub-cluster described in literature have >45% identity, 50% of the genes have >50% identity, or 25% of the genes have >70% identity at the amino acid level, with some variations to account for cluster-type-specific modes of evolution (see **Methods**).

7. BGC alleles and nearest neighbors: An unexpected role for concerted evolution

During an analysis of similarities among several known BGCs, we noted a surprisingly high similarity between the BGCs encoding daptomycin and glycopeptidolipids (GPLs). Although both molecules are lipopeptides, the *Mycobacterium* GPLs are shorter (tetrapeptide vs. tridecapeptide), cell-wall-

associated rather than diffusible, linear rather than cyclic, and from an actinomycete genus that is not closely related to *Streptomyces*.

In our comparison of the rapamycin gene cluster with the BGCs encoding FK520, meridamycin and pladienolide, we discovered that concerted evolution is likely to have had a prominent role in shaping their DNA sequences. In the case of the rapamycin family, recombinations are likely to occur neutrally and have no effect on the structure of the small molecule product (rapamycin, meridamycin and FK520), whereas in other cases, single crossovers within or between gene clusters may dramatically change the modular architecture of a synthase. Near-neutral changes brought about by gene conversion may occur at higher rates for some domains or domain types than for others: in the meridamycin gene clusters, no signs of gene conversion could be observed for KS domains, even though gene conversion manifested itself clearly when comparing the meridamycin clusters with those encoding rapamycin, FK520 and pladienolide. On the contrary, AT domain gene conversion was widespread even between the two meridamycin gene clusters. One might speculate that for these BGCs, gene conversion events get fixed in the population at lower rates for KS domains because not all KS sequences work equally well for different polyketide chain lengths that occur at different points of the assembly-line, so that the changes brought about by a conversion event are less neutral than for AT domains. Mapping of sequence mutations onto the 3D structure of an AT- and KS-containing protein further supports this hypothesis (**Supplementary Figure 17a**), showing widespread sequence variability at almost every position in the AT domains, except for the residues near the substrate binding site (**Supplementary Figure 17b**). Mutations in KS domains, on the other hand, are mostly restricted to the regions in vicinity, and not in the core, of the substrate binding site and the dimerization interface (**Supplementary Figure 17c**), suggesting their importance in influencing substrate selectivity.

We next explored whether concerted evolution is peculiar to the rapamycin family. To our surprise, we could observe the same pattern of BGC-specific branching of domain sequences in many different classes of NRPS and PKS gene clusters, albeit with notable differences between them. For the closely related macrolides erythromycin, oleandomycin and pikromycin, BGC-specific branching appeared to occur for both KS and AT domains, similar to the pattern for rapamycin, FK520, meridamycin and pladienolide. However, for the closely related ansamycin antibiotics macbecin, geldanamycin and herbimycin and the antifungals pimaricin, nystatin and amphotericin, BGC-specific branching occurs for AT domains but not for KS domains. Finally, domains from the trans-AT PKS gene clusters encoding pederin and psymberin do not show any BGC-specific branching at all.

We observed that certain NRPS gene clusters also show signs of concerted evolution: a clear BGC-specific branching pattern pointing to concerted evolution can be seen for the A domains and most of the C domains of the gene clusters encoding the closely related calcium-dependent lipopeptides daptomycin, A54145 and CDA. However, the glycopeptide gene clusters encoding balhimycin, teicoplanin and A40926 showed no such pattern at all: almost all domains cluster in groups corresponding to domains in the same positions in the assembly line.

Collectively, these observations suggest that concerted evolution is a key mechanism driving the evolution of NRPS and PKS gene sequences, but the extent to which it happens depends on functional constraints as well as on the presence of other evolutionary forces acting upon a gene cluster.

We observed several distinct mechanisms of BGC evolution in NRPS and PKS gene clusters (**Figure 8**). Overall, gene clusters encoding glycopeptides, calcium-dependent lipopeptides and macrolides/polyethers appear to be more repetitive, pointing to a history of module duplications or a prominent influence of concerted evolution. The syringopeptin NRPS and mycolactone PKS are outliers: both are likely to have evolved recently by subsequent module duplications and concerted evolution. The PKs psymberin and erythrochelin represent the opposite end of the spectrum: vertical evolution with domain swapping with other gene clusters. Like the trans-AT PKS gene clusters (Nguyen et al. 2008), NRP siderophore gene clusters are relatively non-repetitive, but they seem to have a higher tendency to recruit domains from dissimilar gene clusters. This recruitment appears to be a general feature of NRPS gene clusters as opposed to PKS gene clusters, and might be related to the wider range of possible substrates for NRPSs, which often require BGC-specific sub-pathways for the synthesis of a dedicated monomer (Wilkinson and Micklefield 2009). We observe two more ways in which duplication/divergence and domain swapping can be mixed. First, we often observed gradients of the internal homology p-values from the N- to C-termini of large synthases, suggesting the possibility of N-terminal acquisition of new modules (by duplication or swapping), which would have the effect of extending an intermediate NRP or PK by the addition of a new starter unit. Second, there are many gene clusters showing a 'mixed' mode of evolution, in which one or more of the above mechanisms are combined.

8. Implications for BGC discovery and engineering

Our findings have three main implications for future efforts to discover and engineer BGCs. First, the universe of BGCs is vast, and experimental sampling on it has been biased in terms of phylogeny and of biosynthetic types. Neglected taxa and gene cluster types – particularly RiPPs and saccharides – should be mined more extensively. Second, ClusterFinder identifies a plethora of widely distributed BGCs of unknown function, suggesting that the small molecule products of these clusters should be prioritized for experimental characterization.

Third, efforts to engineer the biosynthesis of unnatural natural products should be based on the 'degrees of freedom' observed in the evolution of specific BGC classes. Our evolutionary analysis of NRPS and PKS gene clusters suggests that concerted evolution has created sets of domains within gene clusters that are highly homologous. These domain sets are more likely to be mutually interoperable than domain sets chosen at random, and might therefore be of great utility in future engineering efforts. In combination with new synthetic biology techniques that enable the rapid assembly of thousands of clusters from a common set of parts (Medema et al. 2011a; Medema et al. 2012; Menzella and Reeves 2007), our results suggest a new approach for re-engaging gene cluster engineering in an manner informed by the mechanisms by which gene clusters have naturally evolved.

Full Methods

Genome information

For all available full genome sequences, gene and Pfam domain annotations were obtained from the JGI-IMG database (Markowitz et al. 2012), version 3.2 (08/17/2010). In the JGI-IMG database, coding regions in prokaryotic genomes are predicted with Glimmer (Delcher et al. 2007), while domains are annotated with HMMER3 (Eddy 2009) using Pfam-A HMM profiles (Punta et al. 2012).

Training set generation

We first searched for all biosynthetic gene clusters in the NBCI Nucleotide database, (http://www.ncbi.nlm.nih.gov/nuccore/) using the search terms "biosynthetic gene cluster", "secondary metabolite", "natural product synthesis", and "biosynthesis". The results set was then manually curated and supplemented by gene clusters identified through a manual search through the scientific literature between 1990 and 2011. These also included known gene clusters from whole genome sequences. Next, we filtered out redundant gene clusters by selecting one random member from each biosynthetic gene cluster family, with a cluster family defined as a connected component in the >0.7 similarity network (the similarities were calculated using a distance metric that adopts sequence similarity of Pfam domains in addition to Pfam domain architecture, as described below in "Biosynthetic gene cluster prediction method: ClusterFinder"). Finally, by comparing the gene cluster entries with the descriptions of the gene clusters in the scientific literature, we manually checked that the biosynthetic gene clusters were full-length, and not deposited to the NCBI Nucleotide database as partial sequences or sequences with large flanking regions not belonging to the biosynthetic gene clusters. This procedure resulted in a set of 677 biosynthetic gene clusters (**Supplementary Table IV**).

Biosynthetic gene cluster prediction method: ClusterFinder

A two-state Hidden Markov Model (HMM) was designed, with one hidden state corresponding to biosynthetic gene clusters (BGC state) and a second hidden state to the rest of the genome (non-BCG state). A vector of observations fed to the HMM is a sequence of Pfam domains in the order in which they appear in the annotated genome. For each of the Pfam domains from the observation vector, the probability of being part of a biosynthetic gene cluster is computed as a posterior probability of the BCG hidden state using the Backward-Forward algorithm (Press et al. 1992). Emission probabilities of Pfam domain types for the BCG state of the HMM were trained by computing Pfam domain frequencies in our set of 677 known biosynthetic gene clusters, using balance training as follows: first, we binned BGCs into 6 classes (NRPS, PKS, terpene, saccharide, ribosomal peptide, and other), based on antiSMASH predictions of biosynthetic classes. Frequencies of all Pfam domains observed in the training set were then calculated for each class separately, and then joined as an average frequency across all 6 classes. At the end, all frequencies were normalized to add up to 1.

To obtain Pfam domain frequencies for the non-BGC state, we first randomly selected one hundred genomes (**Supplementary Table V**), and aligned all their Pfam domain sequences to all Pfam domain sequences from the BGC training set using the blastp algorithm (Camacho et al. 2009). Only hits with an E-value larger than 1e-10 were used to calculate emission probabilities for the non-BGC state. Frequencies of Pfam domains that appear in BGC state but not in the non-BGC state (or *vice versa*) were set to 1% of the frequency of a single observation. The transition probabilities were inferred from manual annotation of biosynthetic gene clusters in the *Streptomyces avermitilis* genome. Around 30% of genes cannot be assigned to any current Pfam family. Consequently, emission probabilities of such cases were set to 1.0 for both states.

After obtaining the biosynthetic gene cluster probabilities for all domains from an input string of Pfam domains, ClusterFinder identifies gene clusters as sets of genes that are at most one gene apart and contain at least one domain with probability of more than 0.2. Finally, ClusterFinder filters out any biosynthetic gene clusters that do not meet any one of the following criteria: (i) having an average BGC probability of >0.4 (as chosen from the second evaluation set), (ii) being longer than the average length of two bacterial genes (2000 bp), and (iii) containing at least one of the class-specific domains (**Supplementary Table VI**). A summary of the ClusterFinder output on all analyzed genomes is given in **Supplementary Table VII**. ClusterFinder was implemented in Python, and is integrated in antiSMASH (Medema et al. 2011b) as well as in the JGI-IMG platform (Markowitz et al. 2012).

ClusterFinder validation

The performance of the biosynthetic gene cluster prediction approach was tested in two ways. First, using ten manually annotated bacterial genomes, we plotted an ROC curve (**Supplementary Figure 18** and **Supplementary Table VIII**), for which we determined an AUC of 0.84. Second, we searched for recently experimentally characterized biosynthetic gene clusters in the literature, and used them to assess the true-positives rate. We found a total of 74 biosynthetic gene clusters not used in our training sets (**Supplementary Table VIII**). 91% of gene clusters were predicted as biosynthetic gene clusters with a median probability (median across all Pfam domains of a given gene cluster) of >0.4. Two out of the six biosynthetic gene clusters with a median ClusterFinder probability lower than 0.4 were found to contain flanking regions not belonging to the actual gene cluster, while the actual gene cluster was detected in the center. Thus, we could conclude that 70 out of the 74 (95%) of the gene clusters had been detected successfully. The remaining four gene clusters from the test set encode two small terpenoid biosynthesis gene clusters, a putative phenolic lipid biosynthesis gene cluster and another putative BGC that did not contain enough Pfam domain similarity with our training set.

When we compared ClusterFinder with antiSMASH (Medema et al. 2011b), antiSMASH proved to be more conservative than ClusterFinder. In a benchmark comparison on the genomes of *Pseudomonas fluorescens* Pf-5, *Streptomyces griseus* IFO13350 and *Salinispora tropica* CNB-440 (**Supplementary Table IX**), antiSMASH detected 62 out of 65 (95.4%) manually annotated secondary metabolite gene clusters, while ClusterFinder detected 59 of these (90.8%). However, ClusterFinder identified 43 (66.2%) unannotated gene clusters that appeared likely to synthesize small molecule metabolites on manual inspection, whereas antiSMASH detected only 5 (7.7%). This highlights the strength of

ClusterFinder in detecting gene clusters irrespective of whether they belong to known or *a priori* specified classes. Among the additional gene clusters detected by ClusterFinder are gene clusters encoding the biosynthesis of, e.g., alginate and lipopolysaccharides, as well as an uncharacterized cluster that was previously predicted to encode a novel type of secondary metabolite (Hassan et al. 2010). In spite of the increased power of ClusterFinder to find unknown gene cluster types, the algorithm has a low rate of clear false positives (4.6%). Another observation from the comparison of the two algorithms was that ClusterFinder algorithm is more accurate at predicting BGC borders (with 14.4±13.3 and 23.1±12.1 incorrectly predicted border genes per BGC for ClusterFinder and antiSMASH, respectively), which aids in calculating a BGC similarity network, since incorrectly predicted flanking regions would result in noisier BGC similarity values.

Annotation of biosynthetic gene clusters

ClusterFinder-detected biosynthetic gene clusters were annotated by antiSMASH to determine their subtypes (e.g., type I polyketide, nonribosomal peptide, terpenoid). The native antiSMASH types were supplemented by a list of profile HMMs for protein domains characteristic of saccharide gene BGCs (**Supplementary Table X**), as well as by fatty acid gene clusters, which could be assigned based on the HMMs that antiSMASH uses in polyketide synthase annotation (Medema et al. 2011b). Gene clusters lacking protein domains characteristic of gene cluster classes included in antiSMASH were binned in a separate class.

Lipopolysaccharide gene clusters were specifically identified by detection of at least one of the following domains: PF01755 (Glycosyltransferase family 25, LPS biosynthesis protein), PF02706 (Chain length determinant protein), PF06176 (Lipopolysaccharide core biosynthesis protein WaaY), PF06293 (Lipopolysaccharide kinase Kdo/WaaP family), PF04390 (Lipopolysaccharide-assembly), PF06835 (Lipopolysaccharide-assembly, LptC-related), PF07507 (WavE lipopolysaccharide synthesis), PF10601 (LITAF-like zinc ribbon domain) and PF04932 (O-Antigen ligase). Capsular polysaccharide gene clusters were specifically identified by detection of at least one of the following domains: PF05704 (Capsular polysaccharide synthesis protein), PF10364 (Putative capsular polysaccharide synthesis protein), PF05159 (Capsule polysaccharide biosynthesis protein), PF09587 (Bacterial capsule synthesis protein PGA_cap). The percentage of saccharide gene clusters not closely related to known saccharide gene clusters was determined by counting the number of BGC in clusters in the BGC network (MCL clustering on >0.5 Lin distance network and I parameter set to 4.0) that do not contain any known gene clusters (see "Gene cluster distance metric and evolutionary network of BGCs" below).

Phylogenetic distribution of BGCs

The phylogenetic distribution of BGCs across the microbial tree of life was plotted using iTOL 2 (Letunic and Bork 2011). The phylogenetic tree used was based on 16S rRNA marker sequences from the corresponding genomes, and was obtained from JGI-IMG (Markowitz et al. 2012). Estimates of within-taxon variation across the tree were calculated using the quadratic entropy index, which allowed us to determine gene cluster diversity at different parts and depths in the phylogeny

(Pavoine, Baguette, Bonsall 2010). Taxonomic classifications of organisms in genera, families, orders, classes and phyla were taken from NCBI Taxonomy (Federhen 2012).

Gene cluster distance metric and evolutionary network of BGCs

To estimate the evolutionary distance between gene clusters, we used a distance metric from Lin et al. (2006) that is a linear combination of two different indices: the Jaccard index and the domain duplication index, with weights of 0.36, and 0.64, respectively. The Goodman-Kruskal γ index, which was included in the original similarity metric with a low weight of 0.01, was omitted, since the conservation of the order between two sets of domains does not appear to have an important effect on the structure of the small molecule product, except in the case of NRPS and PKS gene clusters[36]. Additionally, sequence similarity information was incorporated in the distance metric, by replacing the term $\frac{N_i^P - N_i^Q}{S}$ in the exponent of the domain duplication index with $\frac{N_i^P - N_i^Q - Munkres(D(N_i^P, N_i^Q))}{S}$. Here, *Munkres* represents the Munkres (also known as Hungarian) algorithm (Munkres 1957) for finding of the maximum bipartite matching in a bipartite graph of distances between domains *D* of the type *i* from the two sets to be compared. Due to the large number of domain sequences, the domain distance was defined as the degree of sequence identity. The sequence identities between domains were inferred from multiple sequence alignments constructed using MUSCLE (Edgar 2004) for all the sequences of each Pfam domain. Default parameters were used (i.e., at most 5 iterations), except for domain types with more than 8,000 sequences, for which the number of iterations was set to 3. The distance between all domain pairs of the same type was defined as 1 – sequence identity. The final network was obtained by using a cluster-cluster distance cut-off of 0.5. Visualization was performed using Cytoscape (Smoot et al. 2011).

Comparison of HGT with primary metabolism

In order to be able to compare the rates of horizontal gene transfer (HGT) in BGCs to primary metabolic gene clusters, we modified ClusterFinder to search for histidine and tryptophan biosynthetic operons. Pfam IDs associated with the histidine biosynthesis pathway (PF00475, PF00815, PF01174, PF01502, PF01634, PF04864, PF08029, and PF08645) or with the tryptophan biosynthesis pathway (PF00218, PF00290, PF00465, PF00697, PF01220, PF01264, PF01487, PF04715, and PF08501) were taken from JGI IMG (Markowitz et al. 2012). Trp or His operons were defined as gene clusters containing at least one of these domains with a probability >0.5 and containing at least two of the domains in total. Among 408 organisms searched, 350 histidine and 288 tryptophan biosynthesis operons were identified in 271 and 248 different organisms, respectively. The average number of domains per predicted gene cluster were 2.9 and 3.1, respectively.

To compare gene cluster similarity to the similarity of the organisms harboring these gene clusters, we used the AMPHORA (Wu and Eisen 2008) (August 10th, 2010) dataset, which contains gene sequences from 562 organisms for 30 universally conserved genes. To remove highly similar

genomes, the resulting graph was clustered using default settings in MCL (van Dongen and Abreu-Goodger 2012), and only one member of each cluster was kept for further analyses. This left us with total of 408 organisms. Genes from these organisms were aligned and evolutionary distances were defined using the same MUSCLE-based methods described above for biosynthetic gene clusters. This resulted in 30 distances between each pair of organisms. The distributions of distances of all pairs were tested for normality using a Shapiro-Wilk test. An organism distance map was then built with distances defined as the mean distances of AMPHORA genes.

Phylogenetic profiling

A two-dimensional array with a sequence of consecutive Pfam domains from a given BGC in one dimension and selected organisms (see "Comparison of HGT with primary metabolism") in the other was created for each BGC. The cells in the array consisted of sequence identities between a given domain from a BGC and the most homologous domain (which is also predicted as part of a BGC) from a given organism. Next, for each possible pair of Pfam domains from a given BGC, we calculated a Pearson product-moment correlation coefficient (PMCC) between the two vectors of the best sequence matches. To take rearrangements into account, we hierarchically clustered the rows and columns of the resulting PMCC matrix. Finally, we parsed motifs that are likely to evolve in a correlated fashion by selecting regions on the diagonal of the clustered PMCC matrix with PMCC values >0.5 (repeated by setting the PMCC cutoff to >0.65 and >0.8). Each motif was divided into all possible sub-motifs of sizes between 2 domains and the total number of domains in a motif. Next, we compared the number of (sub)motif occurrences to the number of all possible (sub)motif occurrences in all BGCs that did not pass the PMCC cutoff. Pearson's χ^2 test with Bonferroni correction was applied to test for statistical significance, with the null hypothesis stating that the two values are equal.

Analysis of recent evolution of BGCs

We performed an all-versus-all alignment of nucleotide sequences of known and predicted BGCs using the blastn algorithm. Gene cluster sequences were divided into bins of 1 kb, and then mapped to the most homologous bins from other gene clusters, as well as from the same gene cluster (to test for genomic duplications). 56% of the bins (118,320 out of 212,176) did not map to any homologous regions in the same or other BGCs. Evolutionary events (insertions/deletions, duplications and rearrangements) were detected by a custom-made Python script (available from the authors on request) comparing each alignment of two-gene clusters having at least three matching bins with >70% identity. For indels, a constraint was used that the flanking regions (of size = 2 bins) of each indel breakpoint must be homologous between query and hit gene cluster, and the bin order must be conserved between them.

Comparison of sequence vs. structural similarity of gene clusters and their products

For a given BGC pair, we first calculated sequence identities between all Pfam domain pairs of each Pfam ID, using MUSCLE (Edgar 2004) multiple sequence alignments. A BGC sequence similarity index was defined as the Jaccard index with the size of the intersection represented by the number of Pfam pairs whose sequence identities were higher than the best 10% alignments of all Pfam domains of the same Pfam ID. Taking into account the underlying distributions of sequence identities between all domain sequences prevented misinterpretation of simpler sequence similarity metrics (e.g., an absolute sequence identity threshold) when different evolutionary rates apply to different protein families. We define structural similarity of a given BGC product pair as the Tanimoto coefficient between the two SMILES strings, using FP2 fingerprints from OpenBabel (O'Boyle et al. 2011).

Sub-cluster analysis of known gene clusters

Sub-clusters with known functions from experimentally characterized gene clusters were manually collected from the literature. Sub-cluster sharing between gene clusters from the training set was calculated using blastp. The minimum requirement used to identify a shared sub-cluster between two BGCs was sharing either 75% of the genes with >45% average sequence identity, 50% of the genes with >50% average sequence identity, or 25% of the genes with 70% identity. To account for different modes of sequence evolution of different sub-cluster types, these values were adjusted with the following sub-cluster type-specific cutoffs to obtain a good match between genetic similarity and chemical similarity:

Sub-cluster type	**minimum % of genes, minimum % average sequence identity**	**OR minimum % of genes, minimum % average sequence identity**	**OR minimum % of genes, minimum % average sequence identity**
sugar	75,60	65,65	55,70
nitrosugar	75,65	70,60	50,70
aminocoumarin	75,35	50,40	25,60
aminocyclitol	75,25	50,34	25,60
pyrrole	75,45	50,50	25,70
type I polyketide	75,55	50,65	25,75
type II polyketide	75,35	50,40	25,45
nonribosomal peptide	75,45	50,55	25,65
terpene	75,35	50,40	25,60
MSAS/OSAS	75,40	50,42	25,50
AHBA	75,35	50,40	25,60
enediyne	75,45	50,50	25,70
indolocarbazole	75,45	50,50	25,70
beta-hydroxytyrosine	75,40	50,50	25,60
dehydro-phenylclycine	75,45	50,50	25,70
hydroxy-phenylglycine	75,45	50,50	25,70
3-dimethylallyl-4-hydroxybenzoic_acid	75,45	50,50	25,70
benzoxazolinate	75,45	50,50	25,70
lipid	75,45	50,50	25,70

The final sub-cluster sharing network was drawn with Cytoscape (Smoot et al. 2011).

Multimodular NRPS/PKS gene cluster evolution

To study patterns of evolution in multimodular NRPS and PKS gene clusters, a range of features was calculated describing key characteristics of these gene clusters. The first set of features was based on the topologies of intra-BGC domain similarity networks (with protein domains and sequence similarity representing nodes and edges, respectively) and consisted of the average clustering coefficient, average sequence similarity, graph transitivity, number of 2-4 node cliques, number of connected components in a graph with sequence similarity >50%, and average neighbor degree. We also included as features the number of different Pfam domain types in a BGC, the total number of domains in a BGC, the average number of domains per gene, and the averages and standard errors of best-matching pair sequence identities and internal BGC similarity indices. Two evolutionary indices were also added: the internal similarity index and the vertical evolution index. To obtain the internal similarity index of a gene cluster, we calculated for each of its NRPS/PKS domains the p-value of its closest blastp match inside the gene cluster, given the distribution of the percent identities of all within-gene-cluster blastp hits of all domains of that domain type in the complete set of gene clusters. The internal similarity index was then calculated from these numbers as the mean of all inverse p-values. The same inverse p-values were used for plotting the internal domain similarity across gene clusters. The vertical evolution index of a gene cluster was calculated as the average difference between the p-value of the top 10 percent identities of a domain's blastp hits to all domains from other gene clusters with the p-values of the Lin distances of the gene clusters to the host gene clusters of each of the top 10 hit domains. Consequently, gene clusters with domains with highly similar closest hits to domains in dissimilar gene clusters get a low value, while gene clusters with domains with dissimilar closest hits to domains in similar gene clusters get a high value.

PCA analysis was performed with the aforementioned features as an input. Compound types were assigned using the classifications taken from the primary literature.

Supplementary Material

Supplementary figures and tables can be downloaded from http://rdmy.info/ch10

Chapter 11

Exploiting plug-and-play synthetic biology for drug discovery and production in microorganisms

Published as:

M.H. Medema, R. Breitling, R. Bovenberg, E. Takano (2010) Exploiting plug-and-play synthetic biology for drug discovery and production in microorganisms. *Nature Reviews Microbiology* 9: 131-137.

Abstract

One of the most promising applications of synthetic biology is the biosynthesis of novel drugs from secondary metabolites. Here, we survey a wide range of strategies to control the activity of biosynthetic modules in the cell in space and time, and illustrate how these strategies can be used to design efficient cellular "synthetic production systems." Re-engineered versions of secondary metabolite biosynthetic pathways identified from any genomic sequence can then be inserted into these systems in a plug-and-play fashion.

Introduction

The rapid progress of genome sequencing has revealed thousands of uncharacterized secondary metabolite biosynthetic pathways, many of which are expected to produce novel bioactive compounds such as antitumour drugs, cholesterol-lowering agents, and antibiotics. These pathways are a rich potential source for drug discovery: they constitute a giant evolutionary library of compounds which in contrast to randomly constructed libraries has been highly pre-selected for optimal stability, bioavailability and bioactivity. However, they originate from a diverse range of organisms, including slow-growing soil-bacteria (Lautru et al. 2005), extremophiles (Lentzen and Schwarz 2006), rare plants (Ro et al. 2006), and species from environmental metagenome samples (Brady et al. 2009), many of which are uncultured or unculturable. Moreover, the conditions under which the pathways are active and the compounds are produced at levels sufficient for characterization are often unknown (Gottelt et al. 2010; Scherlach and Hertweck 2009).

New concepts are therefore required for exploiting this richness. Interestingly, in microbial genomes the genes coding for secondary metabolite biosynthetic pathways are usually highly grouped in gene clusters that generally also contain the pathway-specific regulators and transport systems. This allows for their easy identification by *in silico* detection of signature genes or gene domains specific for certain classes of pathways (Li et al. 2009; Starcevic et al. 2008; Weber et al. 2009). However, expressing them to levels sufficient for biochemical characterization is still a large challenge. Current methods to do so include cultivation in different media, heterologous expression, and manipulation of the regulators that control the gene clusters (Zerikly and Challis 2009). Some researchers have given up on the sequence-first approach and are using proteomics to focus on those gene clusters that are already highly expressed naturally (Bumpus et al. 2009). Although all of these strategies enable the functional characterization of single gene clusters, they can be quite time-consuming. We suggest that the tools of synthetic biology (Khalil and Collins 2010; Lu, Khalil, Collins 2009; Purnick and Weiss 2009) allow us to tune newly sequenced pathways in such a way that they can readily be plugged into suitably pre-engineered microbial hosts whose cellular machinery is already optimized for overproduction of compounds synthesized from specific pathway classes, in order to discover novel drugs from these gene clusters in a rapid and high-throughput fashion (**Figure 1**). Subsequently, several intrinsic levels of modularity of the clusters can be exploited in the further engineering of the biosynthetic pathways for optimal production and generation of useful derivatives (**Box 1**).

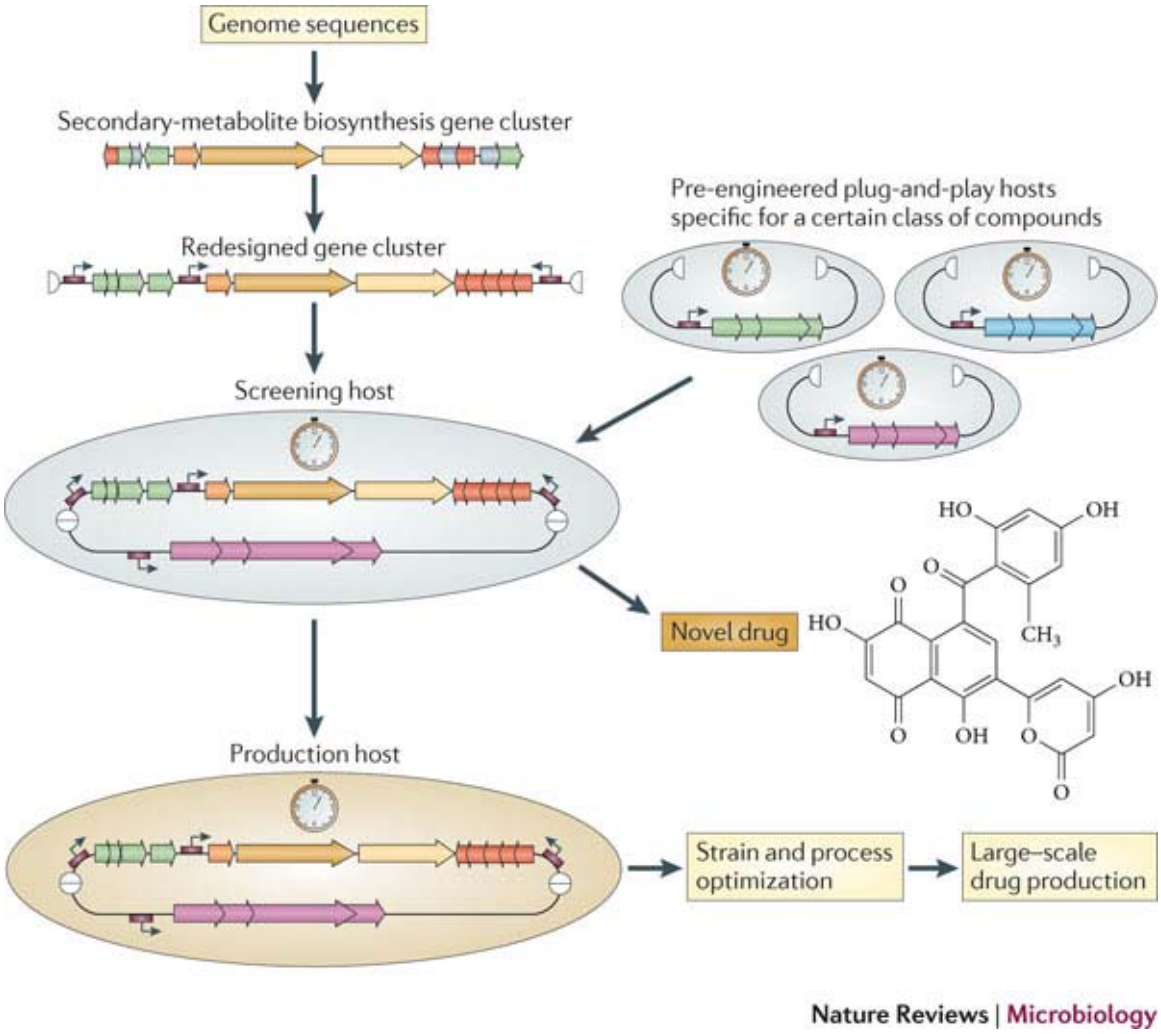

Figure 1: Pipeline for plug-and-play expression of unknown biosynthetic pathways. Overview of the proposed pipeline for plug-and-play execution of secondary metabolite biosynthetic pathways. Gene clusters of interest are selected from the genomic databases and are redesigned for streamlined expression in pre-engineered screening hosts specifically optimized for the broad chemical class of the compound encoded by each pathway. For each screening host, a complementary production host is available to which any compounds with useful bioactivities can be transferred. Further synthetic tuning of these hosts will then lead to efficient production of these new compounds.

The large-scale engineering necessary to make this approach efficient will depend on exerting control over cellular function in both space and time. Here, we delineate the design principles of temporal and spatial control (**Figure 2**) that we believe will be the key to accomplish this new technology. Both temporal and spatial engineering will be discussed at different scales: from allosteric control of enzyme activities to fine-tuning of expression patterns and metabolic programs, and from protein–protein interactions to subcellular organization and microbial communities. Finally, we will discuss how these engineering features can be integrated to design versatile host strains, to enable efficient metabolite screening and production.

Exerting temporal control

When it comes to manipulating biosynthetic pathways for optimal production capacities, the timing of expression of pathway components is important. It is well-known that dynamic regulation of

pathway expression in time generally results in more efficient production than high-level constitutive expression, as in the latter case the continuous metabolic requirements for production conflict with the changing demands for cellular survival and growth. For instance, Farmer and Liao showed that dynamically regulating *Escherichia coli* lycopene production by recruiting the Ntr regulatory system to stimulate production at times of high glycolytic flux resulted in much higher product titers (Farmer and Liao 2000). Synthetic biology can be applied at several scales to achieve optimized exploitation of the metabolic potential of engineered microbes; at the rapid scale of allosteric control of enzyme activities and gene expression; at the intermediate scale of temporal fine-tuning of enzyme expression patterns; and at the long-term scale of synchronizing population activities and executing metabolic programs.

Allosteric control

The most rapid manner in which the cellular machinery can be controlled is directly through metabolite concentrations. This enables a rapid response of enzyme activities to changes in metabolite levels, but also the rapid adaptation of gene expression levels to a specific metabolic situation (Holtz and Keasling 2010). For this purpose, metabolites act as co-factors for regulatory proteins; and as part of a synthetic biology strategy the effector specificity of these regulators can be altered so that they can be employed as switches in different contexts, beyond their natural regulatory role (Galvao and de 2006). For example, the L-arabinose-binding regulator AraC has been engineered to bind D-arabinose instead, through the construction of mutant libraries (Tang, Fazelinia, Cirino 2008). Unfortunately, protein engineering has not advanced far enough to allow for straightforward creation of regulators that bind small molecules that are not so closely related . However, metabolites can also be bound by riboswitches: regulatory RNA elements which reside in the non-coding regions of mRNAs and regulate their translation through changes in their folding pattern induced by the binding of a small molecule. These elements have the great advantage compared to proteins that they can be engineered more easily to bind to a specific small molecule, mainly because RNAs can be synthesized on large scales more straightforwardly. Recently, Dixon et al. mutated an adenine-specific riboswitch cloned in front of a chloramphenicol resistance gene and screened it with ligands derived from a library of small molecules to select for riboswitch–ligand pairs that could activate the resistance (Dixon et al. 2010). In this manner, they were able to engineer riboswitches specific for two rather distantly related molecules, ammeline and azacytosine. In this way it should be possible to create switches that can respond to many different sorts of metabolic input.

For instance, riboswitches could be engineered to recognize the final product of a precursor biosynthesis operon in a drug discovery strain, in order to activate this operon when the cell runs out of precursors. If the intermediates of plugged-in pathways are known, various riboswitches could be engineered to recognize each intermediate and intelligently control the expression of the enzymes governing each step accordingly (**Figure 2**). In such a self-regulating system, toxic accumulation of intermediates would be prevented and metabolic fluxes would be automatically adjusted to changing conditions.

Similarly, when optimizing a gene cluster for industrial drug production, the transcriptional units encoding the transporters could be regulated by a riboswitch that recognizes one of the late

intermediates, in order to have the transporters available just in time to avoid toxic intracellular build-up of the end product.

Timing of enzyme expression

It is energetically wasteful if biosynthetic intermediates are generated that cannot be processed downstream, or if enzymes are being produced by the cell before their substrate becomes available. Enzyme production (and the associated gene transcription) is one of the most costly processes in a microbial cell, in terms of energy and resources, and its optimization has clear evolutionary benefits (Wagner 2007). In the development of screening strains, exquisite tuning of enzyme expression through the engineering of transcriptional units, promoters and ribosome-binding sites (RBSs) is therefore needed to achieve optimally timed fluxes towards the metabolic precursors (**Box 1**). Zaslaver et al. (2004) first reported the phenomenon of 'just-in-time' gene expression in *E. coli* amino acid biosynthesis, in which enzymes are consecutively produced in distinct phases in an order corresponding to their sequence in the pathway. In this manner, wrongly timed superfluous enzyme production is prevented. Similar just-in-time gene expression has been observed in the biosynthesis of exotic polyketide synthase (PKS) starter units and nonproteinogenic amino acid non-ribosomal peptide synthetase (NRPS) precursors (Sattely, Fischbach, Walsh 2008). Such sequential activity patterns can be generated by differential binding affinities of transcription factors to the promoters of the genes involved (Chechik et al. 2008). A library of transcription factor binding sites with different strengths would allow using this coordinated regulation in artificial biosynthetic operons. Combining the construction of the libraries with *in silico* modeling of regulatory outputs based on experimentally determined regulatory responses could enable the predictable design of entire regulatory networks without the need for extensive tuning (Ellis, Wang, Collins 2009). The approach could be extended to temporal ordering of whole operons by adding yet another layer of regulation in which the expression of synthetic pathway-specific regulators acting on these operons is also ordered sequentially.

By using synthetic regulation to optimize the timing of enzyme production, the cell can direct its limited resources toward production of the right enzymes at the right time. This is important, as core biosynthetic enzymes are often rate-limiting for total production, e.g. in cephalosporin biosynthesis (Malmberg and Hu 1992). Although in current industrial practice the production of drugs is continuous after the start-up phase, such just-in-time expression could also be implemented to achieve continuous oscillation of pathway expression in synchrony with other cellular programs (Fung et al. 2005), which can sometimes lead to more efficient flux, as is seen in nature for example in yeast glycolysis (Dano, Madsen, Sorensen 2007).

Population synchronization of metabolic programs

To obtain maximal production of compounds encoded by pathways plugged into a strain it would be desirable to fully exploit the biosynthetic resources of the whole population synchronously. Synchronicity of compound production ensures that the maximal fraction of cells is actually contributing to biosynthesis and avoids the loss of resources that would be the unavoidable

consequence of population heterogeneity. Currently, synchronized gene expression is not always achieved in industrial fermentations. Interestingly, 'synchronized bacterial clocks' have recently been generated in which quorum sensing is applied to synchronize oscillatory gene expression in large populations (Danino et al. 2010). Such a system could also be applied to synchronize the activation of genetic programs in drug-producing bacteria.

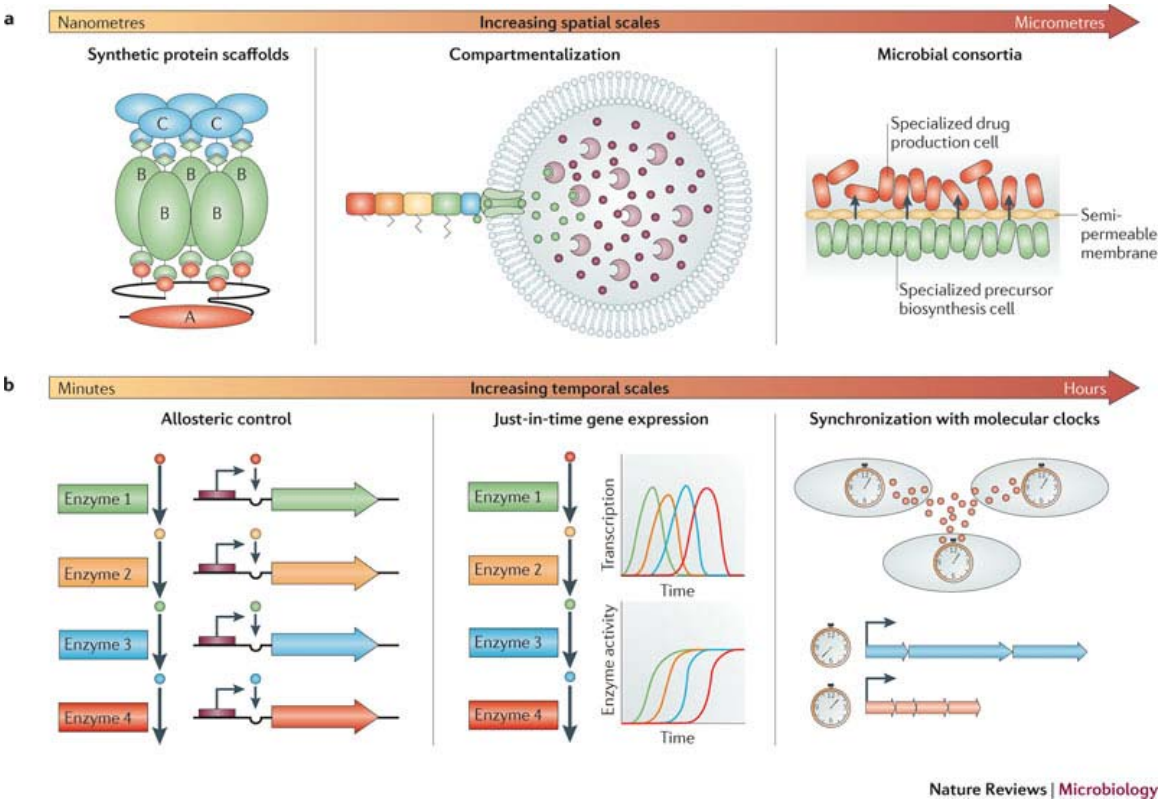

Figure 2: Controlling space and time on different scales in an optimized plug-and-play system. Cellular systems can be controlled at different scales in space (A) and time (B). In space, protein scaffolding can be implemented to physically co-localize enzymes from a certain biosynthetic pathway, and the number of linkers connecting the enzymes can be used to tune stoichiometry. Compartmentalization of pathways can also improve local enzyme and substrate concentrations and thus increase pathway fluxes, but additionally it can make outward transport more efficient and prevent self-toxicity. Microbial consortia can be of great value as they allow the metabolic systems of each cell type to be specifically tuned to a part of the production pathway and hence avoid the need for trade-offs within a single cell type. In time, transcription can be rapidly controlled on the basis of metabolite concentrations through allosteric control, which allows for intelligent tuning of the expression of each enzyme based on the metabolic state of the pathway. One level higher, just-in-time gene expression can ensure that genes are only transcribed when the proteins which they encode are needed, so that protein translation is optimally coupled to the implemented biosynthetic pathway. Finally, synchronization of cell populations can help to synchronize metabolic programs and employ as large a part of the population as possible directly for compound production.

A particular benefit of this synchronization during drug production would be that it would allow the sequential use of distinct metabolic programs: cells could be designed to focus on biomass accumulation in the start-up phase and then switch to a distinct drug production phase. Such metabolic programs also occur in natural producers of secondary metabolites such as *Streptomyces coelicolor*, which has been shown to undergo a major metabolic shift at the transition from exponential to stationary phase (Alam et al. 2010; Nieselt et al. 2010). Yet in a synthetically designed metabolic program, one would not want to wait for depletion of metabolic resources to occur before drug production is started, but switch to this program before the end of the exponential growth

phase, so that full-scale drug production can be coupled to well-adjusted slow growth just high enough to maintain biomass levels.

Exerting spatial control

The optimal engineering of spatial aspects of microbial metabolism is key to the effective design of cellular machinery: the local concentration of enzymes and metabolites determines the rate at which the biosynthetic reactions in a pathway can proceed (Conrado, Varner, DeLisa 2008). The more tightly metabolites are channeled between reaction centers, the more efficiently enzyme resources are used. This is most important when the intermediates and/or end products are hydrophobic and thus limited in their diffusion, but even the spatial arrangement of genes on a bacterial chromosomes appears to be crucial, as mRNA diffusion from the site of transcription in *E. coli* was recently shown to be very limited as well (Montero Llopis et al. 2010). Similar to temporal control, spatial control can be exerted at multiple scales; at the miniature scale of protein complexes; at the intermediate scale of subcellular organization; and at the large scale of entire microbial communities (**Figure 2**).

Scaffolding of enzyme complexes

At the scale of proteins, synthetic protein scaffolds – which assemble proteins into macromolecular complexes – have been successfully employed by Dueber *et al.* to increase local enzyme concentrations and thus to increase the metabolic flux through a biosynthetic pipeline (Dueber et al. 2009). In an *E. coli* host, they linked different numbers of interaction domains from metazoan signalling proteins to biosynthetic enzymes catalyzing subsequent steps of the *Saccharomyces cerevisiae* mevalonate pathway (Dueber et al. 2009). The protein stoichiometry within the scaffolds could be tuned to the catalytic efficiencies of the subsequent enzymes by varying the number of interaction domains that linked them, in order to optimize the flux through the pathway and avoid build-up of intermediates. In the optimal configuration, production titers increased almost eighty-fold, even though enzyme expression was lower than in the native configuration (Dueber et al. 2009). Recently, the scaffolding strategy was also successfully employed to obtain unprecedented high titers of glucaric acid in *E. coli* (Moon et al. 2010). Such protein scaffolds – complemented with protein fusions where beneficial – could well be used to divert and increase metabolic fluxes toward the precursors for certain chemical classes in the development of screening strains. This may prove very useful, as precursor biosynthesis has been shown to quickly become a serious bottleneck in the optimization of product titers. For example, when biosynthesis of the core structure of the important antibiotic erythromycin was optimized using codon-optimized synthetic PKS genes, the synthetic PKS genes were so efficient that they had to be downregulated again to avoid metabolic imbalance of precursor levels, which were too low to constitute an adequate supply to the PKSs (Menzella et al. 2006).

In the secondary metabolite biosynthetic pathways themselves, NRPSs and type I PKSs also function as scaffolds with docking domains that facilitate module–module interactions. Scaffolding has proved to be crucial for efficient pathway flux in such systems: in attempts to construct synthetic

PKSs, only those with functional interaction domain pairs showed significant activity (Menzella et al. 2005; Menzella, Carney, Santi 2007). These assembly lines can also be customized, in order to physically link tailoring enzymes to the domains of the core scaffold, increasing yields of the final product. In such a procedure, one could replace the natural linker domains with linker domains from other protein systems. This strategy could relieve the need for highly product-specific tailoring enzymes, as the co-localization of the enzymes would probably reduce the rates of unwanted side reactions.

Box 1: Building blocks at different levels of modularity

Secondary metabolite biosynthetic pathways display modularity at different levels of organization. Modules at all these levels can be implemented as building blocks which can be recombined in novel ways in the synthetic engineering of secondary metabolism, following the way in which nature engineers these clusters during evolution (Walsh and Fischbach 2010).

The first level of organization of biosynthetic gene clusters is its subdivision into multiple operons or sub-clusters, which are often responsible for the biosynthesis of a distinct part of the end product. In the course of evolution, nature has often recombined such operons in novel ways to generate new compounds. For example, the hybrid antibiotic simocyclinone has likely evolved from fusion of aminocoumarin-producing operons and anthracycline-producing operons (Walsh and Fischbach 2010). This strategy will be mimicked in synthetic engineering.

Going one level down, operons are also modular, consisting of many different genes. The composition of these operons can be altered in various ways. Thus, biosynthetic steps can be removed or novel functionalities can be added. For instance, tailoring enzymes that alter the backbone of a compound can be introduced heterologously (Olano, Mendez, Salas 2010).

At the final level, the core scaffolds of many secondary metabolites are synthesized by polyketide synthases (PKSs) or nonribosomal peptide synthetases (NRPSs), giant multidomain megasynthase enzymes which act as long assembly lines. These contain several modules that catalyze the incorporation of a specific precursor into the growing polyketide or peptide chain. Genomic analysis has revealed that for example the PKSs present in the genome of *Streptomyces avermitilis* have largely evolved through recombination of these modules and the domains that they contain (Jenke-Kodama, Borner, Dittmann 2006). This process has been mimicked in synthetic approaches by rearranging modules from, e.g., erythromycin, rapamycin and pikromycin polyketide synthesis (Menzella, Carney, Santi 2007).

Several tailoring enzymes can also be incorporated into the synthases themselves as internal protein domains. Well-known examples are methyltransferase and oxidation domains. However, examples of transporters, formyltransferases, sulfotransferases, nitroreductases, lanthionine synthetases, and cytochrome p450 oxidoreductase domains fused to these large polypeptides can also be found in genome sequences, opening up possibilities for exploitation by synthetic engineering.

Cellular compartmentalization

An aspect that should not be underestimated is the subcellular organization of enzyme complexes. Straight et al. (2007) have reported that the hybrid NRPS-PKS complexes producing bacillaene in

Bacillus subtilis assemble into an organelle-like megacomplex. The observation that this megacomplex is membrane-associated suggests that this configuration may not only be important for increasing local enzyme concentrations, but also for efficient transport of the product out of the cell (Straight et al. 2007). In fungi, an important part of penicillin biosynthesis is known to take place in the peroxisome (Evers et al. 2004). Also, the biosynthesis and export of aflatoxin in *Aspergillus* have recently been shown to be highly coordinated through accumulation in specific vesicles and subsequent transport to the vacuole (Chanda et al. 2009). As the last two biosynthetic steps leading to the final product are performed within the vesicles, the compartmentalization is likely to prevent self-toxicity, as is known to be the case for many plant secondary metabolites (Sirikantaramas, Yamazaki, Saito 2007). Another important advantage of such dedicated subcellular organizations appears to be facilitation of rapid fluxes through high local concentrations of enzymes and intermediates (Evers et al. 2004). Therefore, compartmentalization should seriously be considered for the construction of efficient screening and production strains.

It is possible to utilize existing compartments of bacteria or fungi and their known protein-targeting systems for this purpose, as has successfully been pioneered by Bayer et al. (2009), who were able to increase production levels of a synthetic methyl halide biosynthesis pathway when the key enzyme was targeted to the vacuole. Yet, this would limit one to species with existing well-characterized targeting systems and could also lead to problems due to competition of the engineered proteins with native organellar proteins. Therefore, a more ambitious aim would be to custom-design bacterial organelles that could be introduced into desirable hosts (Roodbeen and van Hest 2009). The recent discovery and functional dissection of a gene cluster governing the biogenesis of magnetosomes in Magnetospirilli (Murat et al. 2010) offers a possible template for the future biogenesis of such organelles in model organisms like *E. coli*. An orthogonal protein translocation system could then be introduced into these compartments to set them apart entirely for their engineered purpose; in such an effort, it would be crucial to avoid cross-talk with the native secretion systems, which could have a serious negative impact on cellular fitness. Intriguingly, we recently discovered genes encoding NRPSs fused to major facilitator transporters (Medema et al. 2010), opening up the possibility that megasynthase assembly lines can be directly coupled to the membranes of synthetic organelles to allow immediate shuttling of the polyketide or peptide scaffold into the organelle, in which it can then be further modified by a range of tailoring enzymes.

An interesting alternative to membranous organelles is also available: the operon that is responsible for generating proteinaceous microcompartments in *Salmonella* has recently been characterized, and specific N-terminal sequences have been shown to direct proteins to these compartments (Fan et al. 2010). These microcompartments have successfully been expressed heterologously in *E. coli* cells, and GFP could be targeted to them by fusion to an N-terminal targeting sequence (Parsons et al. 2010). The advantage of these compartments is that they strongly co-localize enzymes. At the same time, they allow limited and selective diffusion of small molecules in and out of the compartment, as has been shown for the related carboxysome microcompartments of *Halothiobacillus neapolitanus* (Cai et al. 2009). Interestingly, diffusion selectivity can be altered by mutating the shell proteins (Cai et al. 2009).

Designing microbial consortia

Once novel drugs enter the production phase, there is no reason why the whole production process should be carried out by a single strain. Instead, as the biological route towards the production of a drug usually consists of multiple distinct steps, it deserves further exploration to see whether it could be beneficial to let these be performed by dedicated cell types, which allows the metabolism of each cell type to be specifically tuned to each step instead of forcing a compromise between them within one cell (Brenner, You, Arnold 2008). Microbial communities can be engineered in which different cell types play distinct roles (Weber, Daoud-El Baba, Fussenegger 2007). For example, specialized cells could be constructed to supply biochemical precursors for other specialized cells that convert these precursors into the final product. These cells could then be attached to a surface in a multi-layered biofilm (Stubblefield et al. 2010), or they could be attached to two sides of a semi-permeable membrane which allows diffusion of the intermediates but not of the final product between the two populations (**Figure 2**). Recent experiments have shown that such spatial sequestering of syntrophic co-cultures allows stable growth of both populations as it prevents one cell type being outcompeted by the other (Kim et al. 2008). However, growing two or more types of cells together in a fermentor can be done stably as well, if the cells are engineered to control gene expression of a suicide protein in one of the strains, based on detection of relative cell density of the other strain by a quorum sensing system, such as LuxI/LuxR or LasI/LasR (Balagadde et al. 2008; You et al. 2004). Alternatively, an obligatory cooperative system can be implemented in which each cell type supplies an essential metabolite to the other, as has been done by Shou *et al.*, who constructed a consortium of two *S. cerevisiae* strains that lack either lysine or adenine biosynthesis and supply one another with the essential resource to allow growth (Shou, Ram, Vilar 2007). Such approaches may be the most practical strategies to pioneer for large-scale industrial production and harvesting of compounds on the short-term, as spatial sequestering would require technical innovations beyond standard fermentor growth.

Suitable host strains

All of these control strategies can now be used to design microbial systems specifically tuned for drug discovery, which can subsequently be upgraded to specialized production strains for the large-scale production of each drug. Choosing optimal host strains to serve as a basis for these strategies is of vital importance.

Substantial efforts have focused on transplanting secondary metabolite biosynthetic pathways to fast-growing industrial microorganisms such as *E. coli* or *Saccharomyces cerevisiae* (Boghigian and Pfeifer 2008; Kealey et al. 1998; Pfeifer et al. 2001; Watanabe et al. 2006). Yet, production titers of, for instance, polyketides remained relatively low due to metabolic constraints inherent in the host, as required precursors are present only at low intracellular levels (Boghigian and Pfeifer 2008). The first efforts to redirect metabolic flux toward these precursors have yielded promising results with titers of the polyketide phloroglucinol being increased almost four-fold (Zha et al. 2009).

Unfortunately, transplanting biosynthetic pathways to a novel biological context generally yields unpredictable outcomes, as the metabolic fluxes are not tuned to provide the necessary precursors (Keasling 2008). Bringing parts of primary metabolism under synthetic regulatory control offers a

possible solution. This has recently been practiced in the engineering of biofuel cells, in which the fatty acid degradation pathway was knocked out and genes governing the biosynthetic steps towards fatty esters and fatty alcohols were overexpressed at the same time (Steen et al. 2010). Alternatively, one could attempt to redirect the metabolic fluxes by specific gene knock-outs guided by computational modeling of metabolism (Rocha et al. 2010).

On the short term, current industrial production strains such as actinomycetes optimized using classical strain engineering (Adrio and Demain 2006) may still be quite useful starting points for engineering, given their proven track record. But for a future principled approach to the problem, using genome-minimized production hosts seems to be the most exciting approach, given that they waste only a minimal amount of cellular resources on reactions other than their designed purpose. In addition to the recent successes in reducing existing genomes with a top-down approach (Komatsu et al. 2010; Posfai et al. 2006), the first cell controlled by an entirely synthetic genome has also been constructed recently (Gibson et al. 2008; Gibson et al. 2010; Lartigue et al. 2009). This may soon empower researchers to construct synthetic systems for this type of applications from the ground up.

The near future

In the coming years, technological advances will probably continue at the amazing pace we have witnessed recently. Not only will DNA sequencing advance more and more rapidly, but we will also attain an unprecedented understanding of cellular systems at the levels of RNA, proteins and metabolites. From existing cellular systems we can identify thousands of bioactive compounds, all refined by millions of years of evolution yet mostly uncharacterized in terms of their chemical structure and medicinal activity. At the same time, the costs of DNA synthesis are already dropping rapidly (Carlson 2009) and highly efficient techniques for DNA assembly have become available (Gibson et al. 2009). This inaugurates a new age in which the complete *de novo* synthesis of reengineered gene clusters becomes economically feasible. If we grasp this opportunity now and start to fully utilize all available dimensions of cellular complexity to design bacterial cells specialized for drug discovery and production, this will allow unprecedented access to the riches that nature provides. Achieving this ambitious aim will require the integration of knowledge from many disciplines, from bioorganic chemistry and cell biology to bioinformatics and microbial ecology (Walsh and Fischbach 2010). But when the concepts described in this article are applied in a concerted fashion on a large scale, it may become reality sooner than many a person could have imagined just a few years ago.

Acknowledgements

We thank David Hopwood and Christopher Voigt for constructive comments and suggestions. This work was supported by the Dutch Technology Foundation STW, which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs [STW 10463]. RBr is supported by an NWO-Vidi fellowship, and ET by a Rosalind Franklin Fellowship, University of Groningen.

Chapter 12

Design-based re-engineering of biosynthetic gene clusters: plug-and-play in practice

Parts of this chapter were originally published in:

- M.H. Medema, R. Breitling & E. Takano (2011) Synthetic biology in *Streptomyces* bacteria. *Methods in Enzymology* 497: 485-502.
- H. Frasch, M.H. Medema, E. Takano & R. Breitling (2013) Design-based re-engineering of biosynthetic gene clusters: plug-and-play in practice. *Current Opinion in Biotechnology*, doi: 10.1016/j.copbio.2013.03.006

Abstract

Synthetic biology is revolutionizing the way in which the biosphere is explored for natural products. Through computational genome mining, thousands of biosynthetic gene clusters are being identified in microbial genomes, which constitute a rich source of potential novel pharmaceuticals. New methods are currently being devised to prioritize those gene clusters in terms of their potential for yielding biochemical novelty. High-potential gene clusters from any biological source can then be activated by 'refactoring' their native regulatory machinery, replacing it by synthetic, orthogonal regulation and optimizing enzyme expression to function effectively in an industry-compatible target host. Various parts libraries and assembly technologies have recently been developed that facilitate this process. *Streptomyces* bacteria offer a unique platform to pioneer the refactoring process. Using these organisms allows one to benefit from a long tradition of molecular biology in *Streptomyces*, which provides well-studied model gene clusters as well as a number of specific tools, ranging from cloning vectors to inducible promoters and translational control elements.

Introduction

Today's healthcare would be unimaginable without the employment of natural products. Most of our drugs against bacterial or viral infections, cancers, parasites and other maladies are originally derived from the secondary metabolites of bacteria, fungi or plants. Since the introduction of next-generation sequencing, the amount of genome data for natural product-producing organisms has been increasing rapidly, leading to the identification of an unexpectedly large number of biosynthetic gene clusters: typically up to a few dozen per genome, tens of thousands in total. It has become clear that most of the products of these gene clusters are not produced at detectable levels under laboratory conditions (Bentley et al. 2002; Medema et al. 2010; Nett, Ikeda, Moore 2009); the gene clusters are "silent", "sleeping" or "cryptic". Sequence-guided genome mining and heterologous expression of biosynthesis gene clusters have been successful in awakening some cryptic clusters through knocking out or overexpressing regulatory genes (Aigle and Corre 2012; Gerke et al. 2012; Gottelt et al. 2010), but no high-throughput methodology has been fully developed yet.

Arguably, the biggest hurdle to performing gene cluster characterization quickly and in large numbers has been the fact that gene clusters originate from many different organisms, most of which are difficult or impossible to culture or to manipulate genetically. Recently, we proposed a new strategy to overcome this challenge (Medema et al. 2011a). According to this strategy, gene clusters are extracted from the sequence databases *en masse* and screened in high throughput, subjecting them to a standardized protocol of synthetic biology-guided re-engineering (refactoring) and heterologous expression in pre-optimized plug-and-play hosts (Medema et al. 2011a).

The concept of refactoring derives from software engineering (Fowler and Beck 1999). It was first introduced in synthetic biology by Chan et al. (2005), who partially redesigned phage T7 to make it more amenable to engineering. They reannotated the T7 genome, divided it into its fundamental 'parts', and then simplified the composition and ordering of the parts along the genome. Intriguingly, the process of refactoring can also be applied to entire gene clusters. This was pioneered by Temme et al. (2012), who refactored the entire 23.5-kb/20-gene nitrogen fixation cluster from *Klebsiella*

oxytoca, redesigning the operon structures of the gene cluster and replacing all regulatory elements by synthetic regulatory parts that had been characterized and shown to work in the target host organism, orthogonally from its native regulation. The work of Watanabe et al. (2006) has shown that the replacement of native promoters by synthetic ones can indeed be used to obtain successful heterologous expression of known compounds. If a standardized strategy could be developed to computationally identify those gene clusters from across the tree of life that have the highest potential to yield novel types of chemical structures and biological activities, the refactoring process could be applied to allow high-throughput screening of these gene clusters in optimized hosts. Thus, microbial small molecules could yield a multitude of novel drug leads that could serve to combat bacterial infections, cancer and many other diseases.

Here, we will outline the practical ramifications of the entire process of selecting and redesigning biosynthetic pathways, from the computationally aided selection of candidate pathways to the choice of a suitable host and the specific design choices in refactoring and heterologous expression. Finally, we will provide a more detailed scenario for how these new technologies could be pioneered in *Streptomyces* bacteria.

Prioritizing pathways and gene clusters

A number of computational approaches have recently been developed for effectively mining the bulk of genomic data for biosynthetic gene clusters (Khaldi et al. 2010; Li et al. 2009; Medema et al. 2011b). Yet, the question remains how to prioritize the thousands of gene clusters that result from this exhaustive database search: even if refactoring can be standardized to achieve relatively high throughput, dozens—not thousands—of gene clusters can be characterized with it in the short term. Moreover, randomly trying out gene clusters would be a wasteful process if a lot of information is available to make a good pre-selection.

An effective strategy for prioritization, used in other genomic applications (e.g., (Wu et al. 2009)), is to aim for an unbiased sampling of diversity. Until now, the sampling of natural product diversity has instead been rather biased. Biochemically, certain chemical compound classes (such as macrolides) have been strongly overrepresented in gene cluster characterization attempts. Phylogenetically, certain taxonomic groups (such as actinomycetes) have also been strongly overrepresented. Prioritization attempts could aim to reduce such biases, by sampling more neglected gene cluster families and organisms.

On the other hand, some genomes and certain gene cluster families will be more likely to encode compounds with clinically useful bioactivities, such as antimicrobials: when trying to optimize the sampling of diversity, this diversity should therefore be measured multidimensionally (**Figure 1**). Dimensions of gene cluster diversity could include the sequence diversity of core domains (as used in the recent NaPDoS tool by (Ziemert and Jensen 2012; Ziemert et al. 2012)), the overall architectural and sequence diversity of entire gene clusters (as measured by a distance metric, e.g. (Lin, Zhu, Zhang 2006)), the combinatorial diversity of subclusters encoding specific chemical moieties (Fischbach, Walsh, Clardy 2008), the taxonomic diversity of the organisms encoding the compound,

and the diversity in ecological niches occupied by the source organisms. Based on such data, a mathematical algorithm could then be devised to identify a set of gene clusters that would most strongly complement the already characterized gene clusters. Such a set of high-priority gene clusters would constitute a good starting point for refactoring attempts to explore the uncharted territories of the chemical universe.

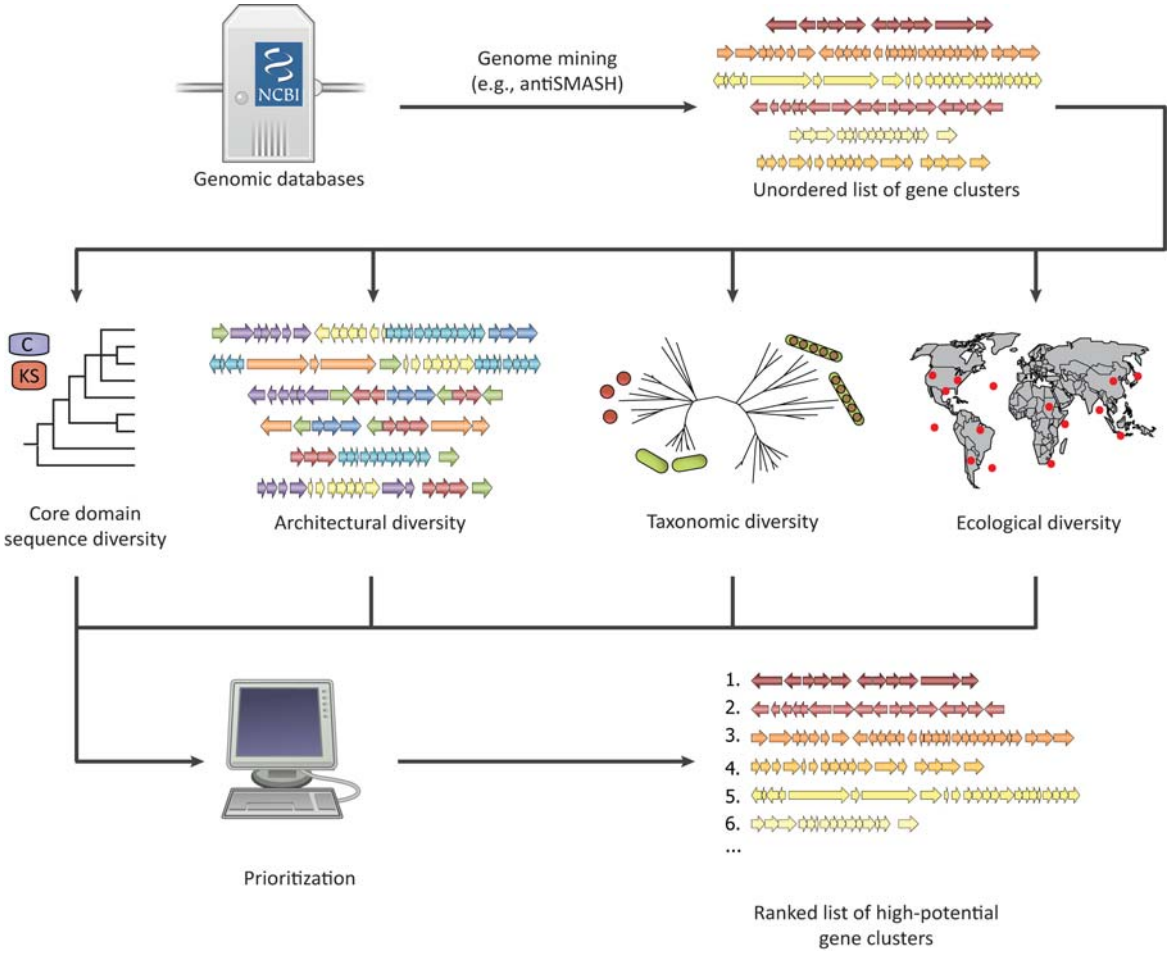

Figure 1: Overview of the process of detecting and prioritizing gene clusters for molecular characterization and refactoring. Gene clusters are first harvested from the genomic databases by software such as antiSMASH (Medema et al. 2011b). Then, for all gene clusters a number of parameters are assembled that allow to survey the diversity of the obtained collection in different dimensions. Finally, a computational algorithm is used to generate a ranked list of prioritized gene clusters that have to be studied to optimize the sampling of diversity at different levels, given the set of gene clusters that have already been characterized before.

Of course, this diversity-enhancing sampling strategy can also be applied in a local fashion, e.g. identifying maximally diverse sets of modifying genes or gene units associated with a specific biosynthetic class of clusters. In the short run, this locally targeted approach might be the more promising, as it increases the probability of hitting on relevant bioactivities by concentrating the search on variations of well-established chemical scaffolds with demonstrated clinical potential, while still expanding it far beyond what is accessible by traditional screening methods.

Finding suitable host organisms

Once a gene cluster has been selected, the first important decision is usually the choice of host organism for overexpression. Naturally, this host will be easy to manipulate genetically, offer a clean background by not producing too many similar compounds itself, and ideally will have a metabolic setup that allows high-volume diversion of fluxes towards the precursors of the expected end compound of the cluster of interest. The advantages and limitations of various well-established industrial hosts, including bacteria such as *Streptomyces* spp., *E. coli*, *Pseudomonas* spp., *Bacillus* spp., and fungi such as *Penicillium* spp. and *Aspergillus* spp., for the heterologous production of polyketides, nonribosomal peptides and isoprenoids has recently been reviewed extensively (Zhang et al. 2011). The currently available host strains are generally good starting points, but are not necessarily optimal: during evolution, organisms are usually not optimized for maximal production titers of secondary metabolites. One strategy to obtain better hosts for a certain class of chemicals is to use organisms that originally produce similar compounds as a starting point: besides the metabolism and metabolite-dependent transcriptional regulation being already tuned to achieve at least moderate levels of compound, this also reduces concerns about self-resistance in the case of antimicrobial activity of the end product, as many antibiotic producers possess redundant resistance mechanisms (Dhote, Gupta, Reynolds 2008). Another strategy uses metabolic modeling to identify those organisms that have an overall metabolic network topology that is best suited for the production of certain compound classes (Zakrzewski et al. 2012). These can then be engineered at the regulatory level to divert their metabolic fluxes in such a way as to actualize this—sometimes hidden—potential.

While engineering host strains derived from species that are not traditionally used in biotechnology may unlock much hidden potential in the long run, a large amount of additional engineering for desirable properties beyond secondary metabolite production will be necessary for such species, e.g. to improve overall growth rate on cheap nutrient sources, to facilitate growth at high density, easy physical handling and genetical manipulation, or to provide necessary biosafety features. Given the costs of this process (and the required approval by regulatory agencies), an attractive alternative approach would be a thorough re-engineering of organisms that are already generally recognized as safe ("GRAS organisms"). Steps in this direction have already been taken, e.g. in the genome-minimization of *Streptomyces* species to reduce background secondary metabolite levels (Gomez-Escribano and Bibb 2011; Komatsu et al. 2010).

Picking up the pieces

When refactoring a biosynthetic gene cluster, most of the genes in the final construct will be the same as in the original gene cluster, but the overall architecture and codon usage will be greatly modified to enable effective heterologous expression (Temme, Zhao, Voigt 2012) (**Figure 2**).

Having a precise annotation of the original sequence (and reliable sequence data, preferably verified by re-sequencing) is crucial to be able to accurately remove all native regulation, such as regulatory genes, promoters, ribosomal binding sites and small RNAs. Changing the codon usage of the open reading frames themselves has the dual function of boosting translational efficiency in the heterologous host and removing unknown regulatory elements that were hidden in the coding

sequence. The new sequence should again be scanned for known internal regulatory elements or undesirable secondary structures that may have been created accidentally (Temme, Zhao, Voigt 2012). Depending on the host, it might also be advantageous to change the start codons: e.g., for streptomycetes it has been shown that genes starting with TTG are better transcribed than genes starting with ATG or GTG (Myronovskyi et al. 2011).

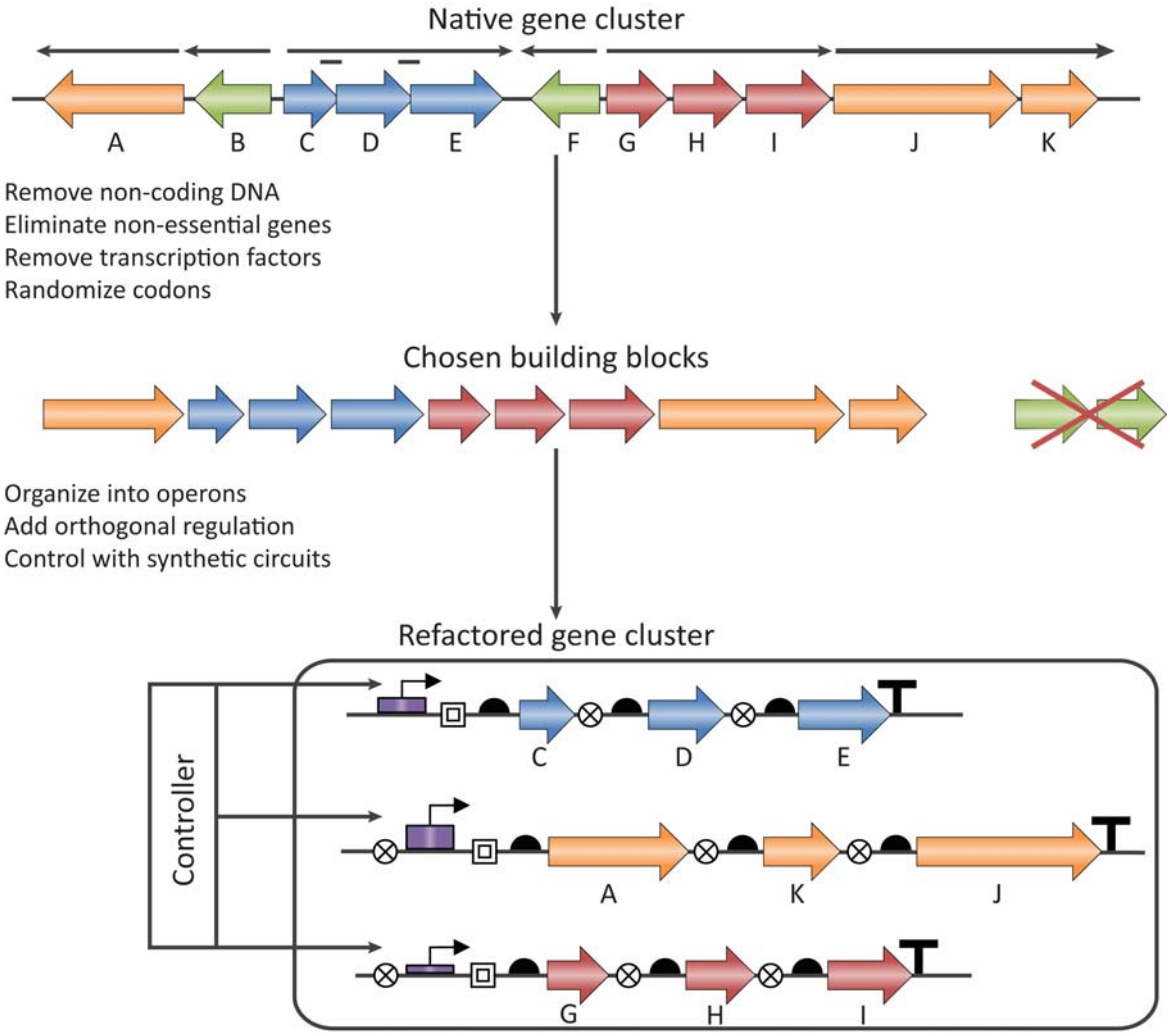

Figure 2: Overview of the refactoring process. At the top, the chosen gene cluster in its native form is shown: arrows above the genes represent transcriptional units, and horizontal bars represent regions were genes overlap each other. The middle part shows how the parts chosen for the refactoring comprise all genes for which a (putative) function in biosynthesis can be assigned. Native regulatory genes are delete/removed. The bottom part shows how the genes are subsequently embedded in a template of standardized parts, with promoters (arrows with purple boxes that signify their strength), ribosome binding sites (filled half circles), insulators (circles with internal crosses) and terminators (dark T's), which is finally assembled and transferred to the chosen host for heterologous expression. Adapted from Temme et al. [10].

The operon structure of a gene cluster can often be simplified during refactoring: genes can be pooled into single operons if they are likely to function in a common process, based on experimental data or evolutionary co-conservation. In the study by Temme et al. (2012), the refactored gene cluster from *Klebsiella oxytoca* was redesigned to form four new operons, each under the control of its own promoter and framed by a terminator sequence. Within an operon, each gene had its own ribosomal binding site and was separated from the neighboring genes by a spacer sequence.

After all native regulation in a refactored gene cluster has been removed, it has to be replaced by new inducible regulatory circuits. The circuit should be orthogonal to the native regulation of the host to avoid cross-talk. Recent breakthroughs and the general principles and strategies of engineering orthogonal regulatory circuits have been reviewed elsewhere (Rao 2012).

For the synthetic pathway to function, the operons need to be expressed in the correct stoichiometry. This can be achieved by choosing promoters with different strengths from a library of characterized promoters. Such a promoter library was created for *Streptomyces coelicoelor*, the model organism of antibiotic-producing actinomycetes, by Rodríguez-García et al., by fusing the *tetO1* and *tetO2* operator sequences to variations of the *permE1* promoter (Rodríguez-García et al. 2005). TetR from the Transposon *Tn*10 is a repressor protein that binds strongly to the *tetO* operator on the DNA and suppresses binding of the RNA polymerase. One promoter from this library, *tcp830*, showed extremely low leaky transcription and very high induction ratios and was recently successfully employed for inducible production of novobiocin (Dangel et al. 2010). There are also different versions of TetR available that bind other DNA-sequences or are induced by compounds other than tetracycline. For many other hosts, similar promoter libraries have been developed as well (Blazeck and Alper 2012).

Recently, Temme et al. (2012) published a complete modular and orthogonal regulation system that utilizes the phage T7 and T3 RNA polymerases. Since these polymerases are in principle highly specific for their cognate promoters, there is little to no cross-talk with other systems in the cell. The authors also provide a library of modified T7 and T3 promoters that can be used to tune gene expression, and showed that their system can be used to regulate two different pathways in *E. coli* without cross-talk (Temme et al. 2012). Another option for setting up orthogonal regulation is through a regulatory circuit from bacterial two-component systems, as pioneered by Whitaker et al. (2012), who incorporated modular scaffolds from eukaryotic signal transduction pathways that enabled redirection of the phosphate from a histidine kinase to different non-cognate response regulators to achieve different outputs.

Besides regulating protein expression at the transcriptional level through the use of promoter libraries, it can also be controlled at the level of translation by varying the ribosomal binding sites (RBSs). RBSs can be engineered with the help of programs like the RBS Calculator (Salis, Mirsky, Voigt 2009) or RBSDesigner (Na and Lee 2010), which aid in constructing synthetic RBSs with a desired translation initiation rate for a given coding sequence.

The strength of RBSs and promoters is context-dependent: different gene or protein expression rates are usually obtained with different upstream or downstream DNA sequences (Lou et al. 2012; Salis, Mirsky, Voigt 2009). Recently, Lou et al. (2012) could show that the rates of transcription can be decoupled from the contextual effect of the junction with the downstream part, by using ribozyme-based insulators between the promoter and the RBS to create uniform 5'-UTR ends of mRNA (Lou et al. 2012). Davis et al. (2011) have also generated a library of insulated promoters for *E. coli*, for which hardly any context effect of the upstream and downstream regions was observed. Such insulators could be utilized to lower the design complexity in refactoring approach by, e.g., using promoters of the same strength for each operon and only introducing variation in the RBSs to arrive at the right protein expression stoichiometry. In general, the correct stoichiometry between the expression of proteins will not be known when refactoring and expressing an unknown gene

cluster, so a library of gene cluster variants with different RBS-gene combinations may have to be created to find those regulatory conditions under which the pathway successfully operates to synthesize its final products at high levels.

To facilitate reliable transcription termination, terminators derived from various phages have been used and shown to function in *E. coli* and *Streptomyces* (Du, Gao, Forster 2009; Du, Villarreal, Forster 2012; Ward et al. 1986). A list of terminators for various hosts can be found in the catalogue of the registry of standardized biological parts. It should be kept in mind that usage of several large class I terminators in close proximity in one construct may lead to instability by recombination (Du, Gao, Forster 2009).

If a refactored pathway or regulatory circuit is not functioning, this can have a multitude of reasons, some of which have been reviewed elsewhere (Cardinale and Arkin 2012). Problems on the regulatory or transcriptional level can be identified by transcriptional analysis and can be addressed by changes in the regulatory system or by changing the order of the genes in a transcriptional unit to create a more stable genetic context (Du, Gao, Forster 2009). Metabolic profiling can identify accumulating metabolites due to bottlenecks in the biosynthetic pathway (Nguyen et al. 2012). The corresponding enzymes can then either be more highly expressed or exchanged with others with higher activity. Tseng et al. (2012) established an optimized pentanol biosynthesis pathway in *E. coli* by dividing it into three modules, each with a defined intermediate, and systematically testing enzymes from a number of different organisms to find ideal combinations for each module (Tseng and Prather 2012).

Assembling the pathway into a synthetic gene cluster

There are various methods available that can be utilized to assemble large DNA constructs both *in vivo* and *in vitro*. These have also been extensively reviewed elsewhere (Ellis, Adie, Baldwin 2011; Ma, Tang, Tian 2012; Zotchev, Sekurova, Katz 2012). However, for the assembly of gene clusters containing large polyketide synthase (PKS) or non-ribosomal peptide synthase (NRPS) genes, which are involved in many of the most interesting secondary metabolite biosynthesis pathways, special challenges have to be overcome. These enzyme complexes can easily be up to 10,000 amino acids in size, comparable to prokaryotic ribosomes (the total clusters often exceeding 100 kb), and are highly repetitive in sequence due to their multimodular domain structure (Fischbach and Walsh 2006). This frequently leads to deletion of parts of the genes by recombination events. It is therefore unlikely that *in vivo* assembly in yeast will be possible for large PKS- or NRPS-genes in the near future. However, recently a (small) PKS cluster has been reassembled in its native state in yeast using the commercial GATEWAY™ technology (Invitrogen, Carlsbad, CA, USA).

Alternatively, there are methods for iterative DNA assembly into the genomes of *Bacillus subtilis* (Itaya et al. 2008; Itaya and Kaneko 2010) and *S. cerevisiae* (Wingler and Cornish 2011). These are the methods of choice for creation of libraries of one pathway, but not for assembly of many different pathways at once.

Therefore, *in vitro* methods for DNA-assembly appear particularly appealing, because here it is possible to assemble all parts of the DNA in one reaction. The methods developed by Gibson and

coworkers are capable of assembling double and single-stranded DNA in reaction times between 45 and 120 min and were used to reconstruct the 500 Mb large genome of *Mycoplasma genitalium* (Gibson et al. 2010). However, depending on the size of the final construct and the cell wall physiology of the selected host organism, the transfer into the host may become a major obstacle.

A final option is complete synthesis of the entire construct. Although this is also challenging for highly repetitive DNA-sequences, the complete synthesis of the erythromycin biosynthetic gene cluster with customized restriction sites for combinatorial biosynthesis has been reported, and the result was shown to be functional in *E. coli* (Kodumal et al. 2004).

A model gene cluster: the actinorhodin biosynthesis cluster from *Streptomyces coelicolor*

Of course the development of the synthetic biology approaches mentioned above needs a concrete starting point to pioneer the engineering in practice. One sensible target for this is a class of antibiotic compounds called polyketides that are synthesized by so-called type II polyketide synthases (PKSs) (Hertweck et al., 2007). Among the polyketides produced by type II PKSs are chemically very diverse bioactive compounds, including tetracycline antibiotics, anthracycline chemotherapeutics (e.g. daunomycin and doxorubicin), angucyclines with a wide range of antibiotic and antitumor activities (e.g. landomycin), and the benzoisochromanequinones, which include the widely studied antibiotic actinorhodin from *S. coelicolor* (**Figure 3**) (Hertweck et al., 2007).

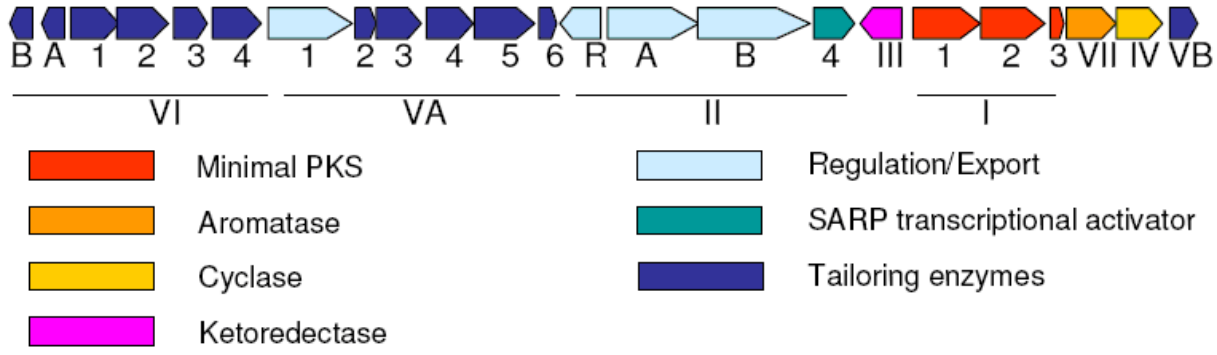

Figure 3: The actinorhodin gene cluster, an example of a well-studied type II PKS polyketide biosynthetic gene cluster. Genes are annotated using different colors as indicated.

The exclusively bacterial type II polyketide synthases are single-domain proteins that form a complex that acts in an iterative fashion to produce the polyketide scaffold that can afterwards be modified by a variety of accessory enzymes (Hertweck et al., 2007). The gene clusters containing type II PKSs are smaller than most other secondary metabolite biosynthetic clusters, making a synthetic approach as outlined above particularly feasible. Moreover, the genetic parts of type II PKS systems, including the post-PKS tailoring reactions, have been shown to be generally interchangeable (Ichinose et al., 2001) and combinable (Hopwood et al., 1985;Khosla and Zawada, 1996;McDaniel et al., 1995) to produce functional compounds. Finally, more than one hundred cryptic gene clusters of this type are currently present in the databases, and this number is still increasing rapidly. Consequently, these gene clusters offer great potential for drug discovery.

Starting from the model organism *S. coelicolor* not only has the advantage of being able to use a wide range of existing molecular tools, but additionally a strain of this species is available in which all four highly active ("non-cryptic") antibiotic biosynthesis gene clusters (actinorhodin: *act;* undecylprodigiosin: *red;* calcium dependent antibiotic: *cda;* coelicolorpolyketide: *cpk*) have been deleted from the chromosome, freeing metabolic resources for the synthetic pathways that are inserted (Gomez-Escribano and Bibb 2011). Alternatively, comprehensively genome-minimized strains would be interesting hosts, such as the recently published genome-minimized *Streptomyces avermitilis* strains (Komatsu et al., 2010) or a plasmid-cured strain of *Streptomyces clavuligerus,* as was recently suggested (Medema et al., 2010).

Practical considerations for synthetic biology in *Streptomyces*

As our understanding of gene regulation in *Streptomyces* is not yet detailed enough to perfectly predict the functioning of all components in an integrated pathway, synthesizing and inserting a whole gene cluster at once is very likely to result in problems that cannot easily be traced to a particular gene. It is, therefore, more promising to use an iterative strategy in which the target gene cluster is subdivided into independent transcriptional units that are individually optimized. After the first design and synthesis of a transcriptional unit, it can be tested for complementation of a deletion mutant of the same genes in the native pathway, by inserting it into the *S. coelicolor* chromosome at the phiC31 or phiBT1 sites (see below). If this step is unsuccessful, the problem can be traced and the design adapted until a successful complementation of pathway functionality is achieved.

Initial proof-of-concept studies can focus on pigmented compounds, such as the abovementioned actinorhodin antibiotic of *S. coelicolor*. In that way, the 'debugging' or 'troubleshooting' process can be aided by photospectrometry to rapidly identify eventual blocks in the biosynthetic pathway, as many intermediates or shunt products produced by knock-out mutants of tailoring genes show absorption spectra distinct from the end product. For instance, in the case of actinorhodin, which has a dark blue color, *actVI-orf1* or *ActVI-orf2* mutants produce a brown pigment (Taguchi et al., 2000), *actVI-orf3* mutants produce a reddish pigment (Taguchi et al., 2000), and *actVA-orf5,6* mutants produce a yellowish brown pigment (Okamoto et al., 2009). A second blue pigment, gamma-actinorhodin, is also known to be produced by the same gene cluster (Bystrykh et al., 1996).

When photospectrometry is not informative, as will be the case for most bioactive compounds targeted in high-throughput genome mining approaches, it will still be possible to predict the structures of most pathway intermediates based on genome annotation (Aoki-Kinoshita and Kanehisa, 2009). In these cases, high-accuracy liquid chromatography mass-spectrometry can be a promising tool for identifying metabolic signatures that characterize bottlenecks at specific steps in the biosynthetic pathway (Kol et al., 2010).

Transcriptional control engineering in *Streptomyces*

While the individual transcriptional units are optimized independently, the synthetic operons can all be controlled by thiostrepton-inducible *tipA* promoters (Takano et al., 1995), and phage fd

bidirectional terminators which are functional both in *E. coli* and *Streptomyces* can be used for transcription termination (Ward et al., 1986) (see below). However, once the complete synthetic gene cluster is assembled, one would most likely need to control the timing and expression rate of some operons separately. For instance, in our example of type II PKS engineering, we might want to control the transcriptional unit encoding the core PKS proteins independently of those encoding the tailoring steps. This would especially be advantageous when genes from cryptic pathways will be inserted in these units later on, which may well need a different mRNA expression stoichiometry compared to the initial model cluster. In this case, at least two promoters with different timing and strength would be required. The use of inducible promoters would have the advantage that one can start at low induction rates to avoid build-up of toxic intermediates, and increase induction later.

Concomitant expression of the cluster-specific secondary metabolite transporters will also be required for toxic compounds. In the wild type gene cluster, expression of the transporter genes is often governed by an intricate system: in our example of the actinorhodin gene cluster, repression of actinorhodin transporter expression by ActR is abolished by binding of ActR to intermediates in the biosynthetic pathway. In this way the transporters will be produced just in time to avoid bacterial suicide (Tahlan et al., 2007; Tahlan et al., 2008; Willems et al., 2008). It is expected that by simultaneously expressing the tailoring genes and the transporters, toxic effects to the cell can be avoided in a similar fashion. Yet, if toxicity problems arise due to lack of transport capacity, it can also be appropriate to insert a strong constitutive promoter, such as *ermE* (see below) in front of the transporter genes.

Translational control engineering in *Streptomyces*

As the translational efficiency of redesigned synthetic genes will be different from those of the wild type genes if the wild type RBSs are used – RBS functionality is context-dependent (Salis et al., 2009) and codon usage also affects translational efficiency – new synthetic RBSs have to be designed to restore the wild type stoichiometry of the enzymes. This requires an accurate estimate of the relative wild type translation rates in each operon. This can be obtained, for example, by fusing the relevant RBS-containing sequence of the wild type gene to a GFP or RFP reporter in a high-copy number plasmid, such as pTONA5 (Hatanaka et al., 2008) or pIJ8630 (de Jong et al., 2009), with a constitutive promoter in front of it; screening for activity is then easily done by measuring the resulting fluorescence. The necessary synthetic RBSs can be identified in the same way, using a library of *Streptomyces* RBSs in the context of each synthetic gene based on oligonucleotides randomized around known RBS sequences. Subsequently, the translation rates of all proteins in the synthetic cluster can be balanced by inserting RBSs from this library that closely match the wild type RBS strength. Using this methodology, RBSs can even be constructed to match translation efficiencies of those unconventional genes in *Streptomyces* that do not have RBSs / UTRs and for which translation starts at the far 5' end of the transcript at the transcription start site itself (Fernandez-Moreno et al., 1994; Strohl, 1992).

Conclusions

For many years, natural products scientists have dreamed of exploiting the modularity of biosynthetic gene clusters to combinatorially recombine them in endless ways to generate enormous libraries of 'unnatural natural products', small molecules unlike those found in nature (Giessen and Marahiel 2012; Sherman 2005; Wong and Khosla 2012). This has thus far only met with mixed results, and substantial technological advances will be needed to make these dreams come true (Khosla, Kapur, Cane 2009). For now, the possibility of re-engineering existing biosynthetic gene clusters that have already been pre-selected by natural selection for biological activity is arguably the most pragmatic way to convert the methods of synthetic biology into tangible pharmaceutical output. The many genetic tools that have been developed for *Streptomyces* and the model gene clusters available from this genus make these organisms an excellent starting point for further developing this technology.

This does not take away the fascinating, more distant possibilities for combinatorial engineering. Besides the obvious swapping of PKS and NRPS modules, subclusters that produce different moieties that are attached to similar core structures could be swapped between different refactored pathways to yield novel structures that are unprecedented in nature. If no combination of a moiety with a similar core structure is known, new developments like the ROSETTA3 algorithm (Richter et al. 2011) may be promising for designing an enzyme with the required transfer activity *de novo.* As our knowledge of biosynthetic pathways and their genetics progresses, so will our ability to engineer them.

Acknowledgements

We thank D. Hopwood for critical reading of a draft of the *Methods in Enzymology* manuscript and K. Ichinose, M. C. M. Smith and H.J. Hong for providing figures and sequence information. MHM is supported by the Dutch Technology Foundation STW, which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs (grant number STW 10463). HJF is supported by a Rosalind Franklin Fellowship to ET at the University of Groningen. RB is supported by an NWO-Vidi fellowship, and ET by a Rosalind Franklin Fellowship, University of Groningen.

Chapter 13

A data-driven metamorphosis: how new technologies will change the face of natural products research

Conclusions and Perspectives

The field of microbial natural products research is in the middle of an extensive transformation: the rapid technological developments seen in genomics, bioinformatics, mass spectrometry and synthetic biology will soon make it possible to discover and characterize dozens or even hundreds of secondary metabolites within the scope of one research project. This thesis has contributed significantly to this development by providing computational methods to automatically identify and analyze biosynthetic gene clusters (BCGs), by mapping all BGCs present in currently sequenced bacterial and archaeal genomes, and by devising synthetic biology strategies to exploit this rich potential. Yet, arguably, the metamorphosis of natural products research is still in the pupal phase at this moment. In the next decade, high-throughput genome mining and synthetic biology re-engineering of BGCs, in combination with the sequencing of hundreds of thousands of microbial genomes, will lead to countless new opportunities. What will the study of secondary metabolite biosynthesis look like in five to ten years, and how should the natural products research community prepare for that?

I predict that four of the most important disciplines that will influence natural products research within the next decade are genomics, microbial ecology, high-throughput experimentation and synthetic biology (**Figure 1**). Through technological advances, these will provide vast amounts of data and novel methodologies for research and engineering. To integrate these successfully, a fifth discipline will be crucial. This discipline is represented by the interface of biology and computer science, and encompasses both bioinformatics and computational (systems) biology. For the sake of simplicity, I will refer to it as bioinformatics from now on.

Below, I discuss the four fundamental disciplines that are key for the future of natural product research, outline which developments can be expected in the coming years, and explain the pivotal role that bioinformatics has for each of these developments.

Microbial genomes by the millions

During the last decade, DNA sequencing technologies have undergone an unexpectedly rapid development, turning the sequencing of a genome from a multi-million euro research project to an almost everyday laboratory exercise. It is not at all unrealistic that within one or two decades the nucleotide sequence databases will contain genome sequences of millions of biological species.

For bacteria alone, 19,379 genome projects have already been registered at the Genomes OnLine Database (Pagani et al. 2012). Large genome sequencing projects such as the Genomic Encyclopedia of Bacteria and Archaea (Wu et al. 2009) and the BGI's 10,000 microbial genomes project are adding to this number continuously. And there is much more left to sequence. According to the World Federation for Culture Collections (http://www.wfcc.info/ccinfo/), international microbial strain collections (Smith 2003) contain more than 1.5 million different bacterial and fungal strains. This of course just covers cultured microbes. Uncultured microbes seem to constitute the vast majority of biodiversity on earth: about half of all identified bacterial phyla even contain exclusively uncultured species (Rappe and Giovannoni 2003), and some reports estimate that as many as 99% of all microbial strains are not readily culturable (Streit and Schmitz 2004). Hence, as soon as single-cell genome sequencing of the uncultured majority of microbes (Lasken 2012) (now already covering 874

genome projects on GOLD) gets up to speed and becomes less expensive, the number of genomes that can be sequenced will become almost unlimited.

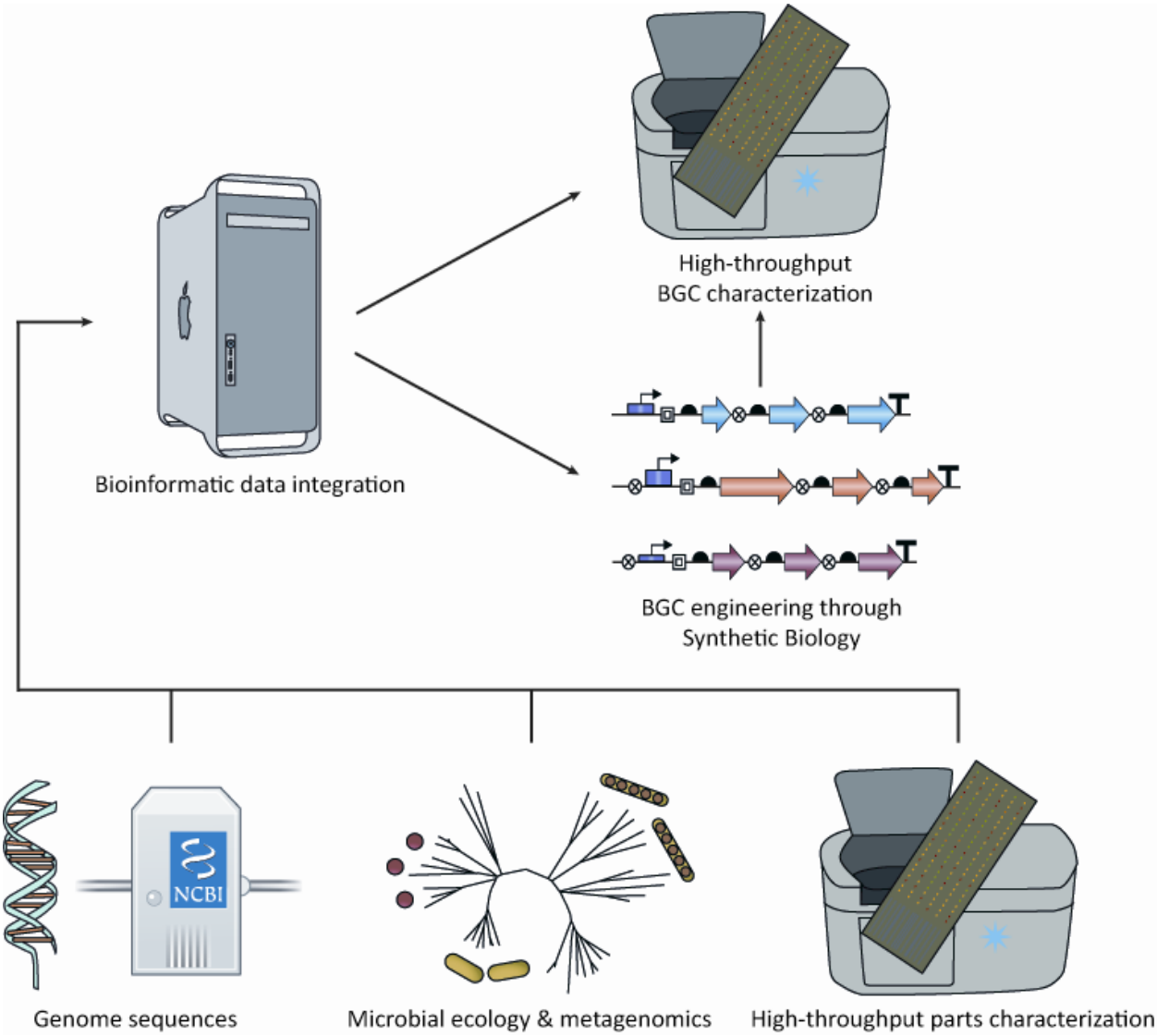

Figure 1. The four key technological driving forces that are likely to revolutionize natural products research are genome sequences and microbial ecology (data input), high-throughput experimental approaches (both data input and BGC characterization) and synthetic biology (BGC characterization and engineering). See also Chapter 12 for a discussion of some of these aspects.

Sequencing one or two million strains within the next decade would mean a hundred- to thousand-fold increase of the amount of data currently available, which is already proving too much to handle with traditional methods in many cases (Beiko 2011). Moreover, new sequencing technologies will yield longer reads (English et al. 2012), making it possible to complete genomes that would otherwise remain fragmented. This will allow better assembly of repetitive BGC sequences such as those encoding polyketide synthases and nonribosomal peptide synthetases. In the end, these technological advances are likely to further drive the ongoing change in the way genome mining for BGCs is done. Finding interesting BGCs will no longer be the issue; the issue will be how to prioritize them (Chapter 12).

With the increase of genome sequences and BGC sequences, it will become more and more challenging to make full use of all this information. When millions of BGCs are available for study, these will first of all need to be ordered in some meaningful way. For example, a database of BGC families and superfamilies (e.g., a 'BGCfam' system, analogous to the popular Pfam database of protein families) could be created. Such a database could make it easy to assign newly identified BGCs to distinct and well-annotated families, with direct links to experimental data and contextual data wherever available. This would also facilitate the easy mining of large numbers of 'alleles' of any given BGC, and arrive at a more thorough understanding of how evolution modifies various classes of BGCs on both the short and the long term. Whole-BGC-based phylogenies (Fischbach, Walsh, Clardy 2008) could allow researchers to distinguish between several structural variants that may exist within a BGC family.

The automatic detection and detailed annotation of subclusters within BGCs (Chapter 10) would also be vital to make sense of all the combinations of these that make up final functional BGC architectures, and to help predict the chemical structures encoded by them. Because of the vast diversity of BGCs, a BGCfam database would still be likely to contain tens of thousands of entries, which would prohibit very detailed annotation. In contrast, the number of subclusters that make up the architectures of these BGC families are likely to be much more limited, which would allow detailed functional characterization and annotation. More detailed analysis of tailoring enzymes such as glycosyltransferases and oxidoreductases, which do not always occur in co-conserved subclusters, would also be vital. The smCOGs constructed in antiSMASH (Chapter 2) could function as a starting point, while this principle should be extended to easily get a clear idea of how a certain gene relates phylogenetically to experimentally characterized family members as well as to family members from experimentally characterized BGCs of BGC families. Using smCOGs, gene interactions could also more easily be predicted using methods similar to those used in STRING (Franceschini et al. 2013), such as co-occurrence within BGCs, physical neighboring on the genome, mentions in the same scientific literature and experimental data of BGC family members.

The torrent of data will only increase in the future. Therefore, adequate standardization, centralization and contextualization of genomic data on BGCs will be vital, so that new BGCs can be straightforwardly linked to information on their BGC family context and associated molecular family, their constituent subclusters and associated chemical moieties, and their encoded tailoring enzymes and associated chemical modifications. Community efforts will be needed to push such databases to a global and all-encompassing level. This has been attempted before by databases such as MapsiDB (Tae, Sohng, Park 2009), but with limited success. To really make this work, the expertise of institutions such as the Genomic Standards Consortium (Field et al. 2011; Yilmaz et al. 2011) will be crucial. Moreover, cooperation with scientific journals should be attempted to make contextual data submission a mandatory requirement for papers reporting the functional characterization of a BGC. If this does not happen, we will soon drown in an enormous mass of underutilized data that is enormously difficult to navigate and exploit.

Understanding chemistry through microbial ecology

Besides regular genome sequencing, metagenomics will continue to offer an increased understanding of microbial communities as a whole. It is increasingly accepted that the vast majority

of microbes cannot be understood adequately in isolation. Instead, microbial communities are the key units that determine the functioning of microbial life on earth. These can be studied using eco-systems biology (Raes and Bork 2008), including the analysis of integrative models based on not only metagenomic but also meta-transcriptomic, meta-proteomic and meta-metabolomic data. The ecological perspective could be of great help to answer fundamental questions about microbial secondary metabolism. For example, Cordero et al. (2012) recently showed that antibiotic production and resistance often functions as means of warfare *between* socially cohesive populations of bacteria composed of various species, instead of between species *within* a population. When metagenomics datasets become available for large numbers of representative environments and conditions, the compositional dynamics of microbial communities and the roles of secondary metabolites therein can be reconstructed on large scales (both spatial and temporal).

This might even make it possible to start answering exciting questions such as: Can we predict the occurrence of certain BGC classes in microbial communities based on the strain composition of a each community? Can we predict in which environments and communities we are most likely to find BGCs encoding molecules with a certain molecular function, such as antimicrobials active against a certain range of bacteria? Can we predict whether certain BGCs are expressed based on the co-occurrence of their 'host' microbe with certain other microbes? Achieving a better understanding of the natural roles of secondary metabolites (Davies and Ryan 2012) will be crucial to adequately answer such questions. Moreover, it will be important to develop computational methodologies to dissect phylogenetic correlations from lifestyle/community/environment correlations that go beyond these. Finally, just as for the chemical attributes of BGCs, the ecological and environmental contextual data on BGCs should also be standardized if any meaningful information is to be gathered from it.

Just like BGCs can be clustered into BGC families and molecules can be clustered into 'molecular families', ecosystems could be clustered into 'ecosystem families.' This will become more and more useful when more metagenomic data are generated. Some ecosystem families might then turn out to be specifically likely to host certain BGC families. When BGC families are annotated further in terms of their molecular functions, it might become possible (e.g., by machine learning techniques) to predict the presence of molecules with specific biological activities based on the community structure within the ecosystem. Importantly, such efforts would require a thorough integration of chemical, genomic, phylogenetic and ecological data.

Perhaps one will also be able to approach the issue from the other side, starting at the strain instead of at the ecosystem: it might be more revealing to know for each strain in which communities/environments it is found at certain minimal frequencies. Currently the amount of available metagenomics data is insufficient to acquire the necessary sampling depth to identify statistically significant patterns. Yet even now, it might already be possible to extract a lot of relevant information about a bacterium's ecological role or niche from its genome sequence, as the presence of genes encoding certain enzymes (the 'genomic environment') indicates how the organism makes a living. An organism's genome functions as a logbook of its ecological history and hence may provide information that will help predict the molecular activities of its encoded secondary metabolites. Naturally, this information can be complemented with data on BGC families and BGC subcluster architectures to fuse together into a knowledge base that will allow more reliable predictions.

This way, scientists may no longer have to screen thousands of different strains from all kinds of different ecosystems for molecules with a desired activity, because they will know exactly where to look to have the best chance of finding potential new pharmaceuticals.

High-throughput experimentation paves the way towards comprehensive insight

Recently, the first approaches have been published to characterize BGCs of certain biosynthetic classes in high throughput. Kersten et al. (2011) developed a method to profile organisms for the presence of compounds in the mass range corresponding to peptide natural products and then to use tandem mass spectrometry to generate amino acid sequence tags that reveal the identity of these molecules. This allows the detected peptides to be linked straightforwardly to ribosomal peptide- or nonribosomal peptide-encoding BGCs. Pep2Path (Chapter 5) even automates the latter process entirely. In a few years, it may thus be possible to rapidly sequence the genomes and generate mass spectrometric profiles for tens or even hundreds of strains of interest, and then computationally connect these datasets to automatically generate a list of novel compound masses with their associated candidate BCGs. Aided by automatic dereplication (Ibrahim et al. 2012), this will then quickly result in a list with BGCs that encode truly novel secondary metabolites, together with their associated masses and (part of) their amino acid sequences. With this information, the chemical elucidation of the compound will be much more straightforward, whereas new techniques such as imaging mass-spectrometry may aid in determining the molecular functions of the most promising compounds (Watrous and Dorrestein 2011). Although these technologies have up till now focused on peptides, it is in principle possible to extend them to other types of molecules (such as polyketides, terpenoids and saccharides). These approaches might also be used to link the thousands of chemical structures of secondary metabolites that have been determined in recent decades either to their corresponding BGCs in their source organisms or to homologous BGC alleles from other organisms, to make it easier to manipulate and engineer them.

To increase the power of these approaches, advances are also needed in the automated matching of genes to the chemical structures of the final compounds. For nonribosomal peptides, for example, more training data is needed to improve the accuracy of substrate specificity prediction algorithms, especially for fungal NRPS modules (Rausch et al. 2005; Röttig et al. 2011). This might be pursued by high-throughput expression of NRPS domains synthesized based on genomic information, and subsequent large-scale ATP/PPi exchange assays. In a similar fashion, tailoring enzymes such as glycosyltransferases (Erb et al. 2009; Luzhetskyy et al. 2008) could be characterized in high-throughput using carbohydrate arrays coupled to mass spectrometry (Ban et al. 2012). To increase the accuracy of predictions of the order of NRPS and PKS enzymes in their assembly lines based on their DNA sequence, PKS docking domains and NRPS COM domains (Weissman and Müller 2008) could be synthesized in high throughput to test large numbers of combinations of these using protein–protein interaction assays.

All such forms of high-throughput experimentation yield large amounts of data. Machine-learning approaches such as support-vector machines (Murty and Devi 2011) or random forests (Touw et al. 2012) could make it possible to employ these datasets for the development of relevant algorithms that allow direct predictions of chemistry from the DNA sequences of BGCs.

Bypassing the complexity of regulation by synthetic biology

Many of the high-throughput approaches discussed above depend on the ability to produce synthetic DNA, but the discipline of synthetic biology offers much more. A large problem in awakening and characterizing BGCs is the staggering complexity of their transcriptional and posttranscriptional regulation (Liu et al. 2013). In fact, most BGCs from cultured organisms are silent under most typical laboratory circumstances. Sometimes, the regulation can be modified through targeted engineering, but this is a tedious exercise that often first requires unraveling much of the regulation that is in place. For BGCs from unculturable and exotic organisms, the problem is even bigger: heterologous expression of these clusters without modification will often even be a useless exercise, as there is a large probability that the promoters and ribosomal binding sites (RBSs) will not function in the new host. Even promising high-throughput approaches such as mass spectrometry-guided peptidogenomics only circumvent this problem without solving it: BGCs from cultured organisms that are naturally silent will remain hidden, and BGCs from unculturable organisms cannot be modified to deduce the functions of the various enzymes. Fortunately, as discussed in Chapters 11 and 12, synthetic biology offers an approach to bypass the native regulation altogether, and replace all native regulatory elements such as promoters and RBSs with synthetic versions. Thus, exotic and 'unreadable' BGCs could be quickly and efficiently 'translated' into BGCs that are readable by host organisms that are easy to work with.

As argued in Chapter 7, synthetic biology has the potential to do even more than re-engineering existing pathways: bottom-up, design-based engineering of new BGCs from scratch would be a long-term goal that would enable researchers to generate incredible amounts of chemical diversity. Two things will be essential for such strategies to produce meaningful results.

First, biosynthetic parts will have to be standardized. Sub-clusters for the biosynthesis of precursor molecules (Wohlleben et al. 2012) need to be catalogued in terms of their function and genomic diversity, and both group transfer enzymes (such as glycosyltransferases and methyltransferases) and modules from multimodular PKS/NRPS enzymes need to be characterized in large numbers and in enough detail. These cataloguing efforts should be linked to the standardization of the BGC contextual data mentioned earlier, so that a firm knowledge base is built up that can function as the basic toolbox for engineering.

Second, we need to be able to understand the 'rules' that determine when sub-clusters and genes encoding group transfer and redox enzymes can work together to build composite molecules. On the one hand, we may learn from nature here: the constraints within which BGCs and their encoded molecules evolve in nature are likely to reflect functional constraints that determine how easily certain enzymes can work together (Chapter 10). Bioinformatically pinpointing the evolutionarily most flexible parts (e.g., promiscuous glycosyltransferases, PKS modules that combine easily with other modules, etc.) or combinations of parts that are often encountered together in nature could be an interesting strategy for engineering libraries of parts suitable for engineering. On the other hand, trying to engineer BGCs with entirely 'unnatural' combinations of sub-clusters and enzymes would also make it possible to systematically discover which types of hybrids are nonetheless

biochemically feasible. For example, one could envision the engineering of hybrids between ribosomal and nonribosomal peptide biosynthesis pathways as a real possibility.

In addition to the enzymatic perspective, successful BGC engineering will also have to take into account genomic and metabolic perspectives. In terms of genome engineering, it is interesting to explore the functional consequences of the genomic structuring of BGCs. It is known that bacterial genome architecture has profound effects on the ability of, e.g., transcription factors to interact with regulatory elements: because of the low diffusion distance from the point of translation (which is co-transcriptional), they bind much more easily to regulatory regions that are in close proximity of the genes encoding them (Montero Llopis et al. 2010). It might be the case that in a crowded bacterial cell (Elowitz et al. 1999), enzymes encoded by adjacent genes are also significantly more likely to interact. Hence, such enzymes may gain a superficially higher specificity for each other's products simply because of local concentration effects (Minton and Rivas 2011). Such speculation is all the more interesting because in fungi, where translation is not co-transcriptional and protein diffusion rates are higher (Mika and Poolman 2011), BGCs consisting of multiple sub-clusters are very rare, while they are quite abundant in bacteria (Chapter 10). Perhaps the physical properties of bacterial cells promote more progressive modes of BGC evolution, by giving the enzymes encoded by nearby genes a 'head start' through the minimal functionality already provided by their close physical association. At the same time, this might help explain why biosynthetic genes co-occur together in gene clusters in the first place, as more general evolutionary models such as the selfish operon model do not seem to have adequate explanatory power (Martin and McInerney 2009). In BGC engineering, knowledge about genome architecture could be exploited by integrating BGCs at optimal positions within the host's chromosome and by designing BGCs in such a way that interacting enzymes are encoded by adjacent genes.

Finally, bioinformatics will also play a key role in informing synthetic biology designs by systems biology modeling. As indicated in Chapter 6, modeling can aid in selecting suitable host organisms for heterologous expression of BGCs, but it can also help in engineering such organisms to support certain precursors needed for the production of various classes of molecules. And as I discuss in Chapter 7, metabolic and thermodynamic modeling could even be used to find new and efficient metabolic pathways that are not known to occur in nature.

Getting ready for the metamorphosis

The coming years will reveal how ready the natural products field is for the revolutionary potential of these various new technologies and approaches. Fully exploiting them will require the extensive modernization of methods as well as community efforts to standardize data on BGCs and their context. Serious investments in infrastructure for microbial cultures and databases, as well as in computational algorithms and experimental synthetic biology engineering tools, are likely to allow the field to undergo a radical transformation into a data-centered, engineering-based science. This will open up unprecedented opportunities to understand the full natural diversity of biosynthetic genes, enzymes and molecules throughout the biosphere and exploit it for the good of humanity. These are exciting times!

List of publications

Submitted for publication:

1. Cimermancic P*, Medema MH*, Wieland-Brown LC, Mavrommatis K, Pati A, Godfrey PA, Koehrsen M, Clardy J, Birren BW, Breitling R, Takano E, Sali A, Fischbach MA (2013) Insights into secondary metabolism from a global analysis of biosynthetic gene clusters. In revision.

Published during PhD Thesis work:

2. Nguyen DD, Wu CH, Moree WJ, Lamsa A, Medema MH, Zhao X, Gavilan RG, Aparicio M, Atencio L, Jackson C, Ballesteros J, Sanchez J, Watrous JD, Phelan VV, van de Wiel C, Kersten RD, Mehnaz S, de Mot R, Shank EA, Charusanti P, Nagarajan H, Duggan BM, Moore BS, Bandeira N, Palsson BØ, Pogliano K, Gutiérrez M, Dorrestein PC (2013) MS/MS networking guided analysis of molecule and gene cluster families. **Proceedings of the National Academy of Sciences USA**, accepted for publication.
3. Blin K*, Medema MH*, Kazempour D, Fischbach MA, Breitling R, Takano E, Weber T (2013) antiSMASH 2.0 – a versatile platform for genome mining of secondary metabolite producers. **Nucleic Acids Research** 41: W204-W212.
4. Frasch H, Medema MH, Takano E, Breitling R (2012) Design-based re-engineering of biosynthetic gene clusters: plug-and-play in practice. **Current Opinion in Biotechnology**, doi:10.1016/j.copbio.2013.03.006.
5. Medema MH, Takano E, Breitling R (2012) Detecting sequence homology at the gene cluster level with MultiGeneBlast. **Molecular Biology and Evolution** 30: 1218-1223.
6. Tobias NJ, Doig KD, Medema MH, Chen H, Haring V, Moore R, Seemann T, Stinear TP (2013) Complete genome sequence of the frog pathogen *Mycobacterium ulcerans* ecovar Liflandii. **Journal of Bacteriology** 195: 556-564.
7. Zakrzewski P*, Medema MH*, Gevorgyan A, Kierzek AM, Takano E, Breitling R (2012) MultiMetEval: comparative and multi-objective analysis of genome-scale metabolic models. **PLoS ONE** 7: e51511.
8. Fedorova ND, Moktali V, Medema MH (2012) Bioinformatics approaches and software for detection of secondary metabolic gene clusters. **Methods in Molecular Biology** 944: 23-45.
9. Nguyen QT, Merlo EM, Medema MH, Jankevics A, Breitling R, Takano E (2012) Metabolomics methods for the synthetic biology of secondary metabolism. **FEBS Letters** 586:2177-2183.
10. Medema MH, van Raaphorst R, Takano E, Breitling R (2012) Computational tools for the synthetic design of biochemical pathways. **Nature Reviews Microbiology** 10: 191-202.
11. Medema MH*, Alam MT*, Breitling R, Takano E (2011) The future of industrial antibiotic production: from random mutagenesis to synthetic biology. **Bioengineered Bugs** 2(4): 230-233.
12. Alam MT, Medema MH, Takano E, Breitling R (2011) Comparative genome-scale metabolic modeling of actinomycetes: the topology of essential core metabolism. **FEBS Letters** 585(14): 2389-2394.
13. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. **Nucleic Acids Research** 39: W339-W346.

14. Röttig M, Medema MH, Blin K, Rausch C, Weber T, Kohlbacher O (2011) NRPSPredictor2 – a web server for predicting NRPS adenylation domain specificity. **Nucleic Acids Research** 39: W362-W367.
15. Medema MH, Breitling R, Takano E (2011) Synthetic biology in *Streptomyces* bacteria. **Methods in Enzymology** 497: 485-502.
16. Medema MH*, Alam MT*, Heijne WH, van den Berg MA, Müller U, Trefzer A, Bovenberg RA, Breitling R, Takano E (2011) Genome-wide gene expression changes in an industrial clavulanic acid overproduction strain of *Streptomyces clavuligerus*. **Microbial Biotechnology** 4(2): 300-305.
17. Medema MH, Breitling R, Bovenberg RA, Takano E (2011) Exploiting plug-and-play synthetic biology for drug discovery and production on microorganisms. **Nature Reviews Microbiology** 9(2): 131-137.
18. Medema MH, Trefzer A, Kovalchuk A, van den Berg M, Müller U, Heijne W, Wu L, Alam MT, Ronning CM, Nierman WC, Bovenberg RA, Breitling R, and Takano E (2010) The sequence of a 1.8-Mb bacterial linear plasmid reveals a rich evolutionary reservoir of secondary metabolic pathways. **Genome Biology and Evolution** 2: 212-224.

Previously published (during B.Sc. / M.Sc. studies):

19. Medema MH*, Zhou M*, van Hijum SAFT, Wessels HJTC, Gloerich J, Siezen RJ, Strous M (2010) A predicted physicochemically distinct sub-proteome associated with the intracellular organelle of the anammox bacterium *Kuenenia stuttgartiensis*. **BMC Genomics** 11: 299.
20. Van Hijum SAFT*, Medema MH*, Kuipers OP (2009) Mechanisms and evolution of control logic in prokaryotic transcriptional regulation. **Microbiology and Molecular Biology Reviews** 73(3): 481-509.
21. Ettwig KF, Shima S, van de Pas-Schoonen KT, Kahnt J, Medema MH, op den Camp HJM, Jetten MSM, Strous M (2009) Denitrifying bacteria oxidize methane in the absence of archaea. **Environmental Microbiology** 75(11): 3656-3662.

*Equal contribution

References

Adrio JL, Demain AL. 2006. Genetic improvement of processes yielding microbial products. FEMS Microbiol. Rev. 30:187-214.

Ahlert J, Shepard E, Lomovskaya N, Zazopoulos E, Staffa A, Bachmann BO, Huang K, Fonstein L, Czisny A, Whitwam RE et al. 2002. The calicheamicin gene cluster and its iterative type I enediyne PKS. Science 297:1173-1176.

Aigle B, Corre C. 2012. Waking up *Streptomyces* secondary metabolism by constitutive expression of activators or genetic disruption of repressors. Methods Enzymol. 517:343-366.

Ajikumar PK, Xiao WH, Tyo KE, Wang Y, Simeon F, Leonard E, Mucha O, Phon TH, Pfeifer B, Stephanopoulos G. 2010. Isoprenoid pathway optimization for taxol precursor overproduction in *Escherichia coli*. Science 330:70-74.

Akopiants K, Florova G, Li C, Reynolds KA. 2006. Multiple pathways for acetate assimilation in *Streptomyces cinnamonensis*. J. Ind. Microbiol. Biotechnol. 33:141-150.

Alam MT, Medema MH, Takano E, Breitling R. 2011. Comparative genome-scale metabolic modeling of actinomycetes: The topology of essential core metabolism. FEBS Lett. 585:2389-2394.

Alam MT, Merlo ME, Hodgson DA, Wellington EM, Takano E, Breitling R. 2010. Metabolic modeling and analysis of the metabolic switch in *Streptomyces coelicolor*. BMC Genomics 11:202.

Alexander DC, Jensen SE. 1998. Investigation of the *Streptomyces clavuligerus* cephamycin C gene cluster and its regulation by the CcaR protein. J. Bacteriol. 180:4068-4079.

Alexander DC, Rock J, He X, Brian P, Miao V, Baltz RH. 2010. Development of a genetic system for combinatorial biosynthesis of lipopeptides in *Streptomyces fradiae* and heterologous expression of the A54145 biosynthesis gene cluster. Appl. Environ. Microbiol. 76:6877-6887.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403-410.

An W, Chin JW. 2009. Synthesis of orthogonal transcription-translation networks. Proc. Natl. Acad. Sci. U. S. A. 106:8477-8482.

Anand S, Prasad MV, Yadav G, Kumar N, Shehara J, Ansari MZ, Mohanty D. 2010. SBSPKS: Structure based sequence analysis of polyketide synthases. Nucleic Acids Res. 38:W487-W496 .

Ansari MZ, Sharma J, Gokhale RS, Mohanty D. 2008. *In silico* analysis of methyltransferase domains involved in biosynthesis of secondary metabolites. BMC Bioinformatics 9:454.

Arnison PG, Bibb MJ, Bierbaum G, Bowers AA, Bugni TS, Bulaj G, Camarero JA, Campopiano DJ, Challis GL, Clardy J et al. 2013. Ribosomally synthesized and post-translationally modified peptide natural products: Overview and recommendations for a universal nomenclature. Nat. Prod. Rep. 30:108-160.

Asadollahi MA, Maury J, Patil KR, Schalk M, Clark A, Nielsen J. 2009. Enhancing sesquiterpene production in *Saccharomyces cerevisiae* through *in silico* driven metabolic engineering. Metab. Eng. 11:328-334.

Atsumi S, Hanai T, Liao JC. 2008. Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. Nature 451:86-89.

Bachmann BO. 2010. Biosynthesis: Is it time to go retro? Nat. Chem. Biol. 6:390-393.

Bachmann BO, Ravel J. 2009. Methods for *in silico* prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. Methods Enzymol. 458:181-217.

Balagadde FK, Song H, Ozaki J, Collins CH, Barnet M, Arnold FH, Quake SR, You L. 2008. A synthetic *Escherichia coli* predator-prey ecosystem. Mol. Syst Biol. 4:187.

Baltz RH. 2008. Renaissance in antibacterial discovery from actinomycetes. Curr. Opin. Pharmacol. 8:557-563.

Ban L, Pettit N, Li L, Stuparu AD, Cai L, Chen W, Guan W, Han W, Wang PG, Mrksich M. 2012. Discovery of glycosyltransferases using carbohydrate arrays and mass spectrometry. Nat. Chem. Biol. 8:769-773.

Bao K, Cohen SN. 2003. Recruitment of terminal protein to the ends of *Streptomyces* linear plasmids and chromosomes by a novel telomere-binding protein essential for linear DNA replication. Genes Dev. 17:774-785.

Bao K, Cohen SN. 2001. Terminal proteins essential for the replication of linear plasmids and chromosomes in *Streptomyces*. Genes Dev. 15:1518-1527.

Barabasi AL, Oltvai ZN. 2004. Network biology: Understanding the cell's functional organization. Nat. Rev. Genet. 5:101-113.

Barna JC, Williams DH. 1984. The structure and mode of action of glycopeptide antibiotics of the vancomycin group. Annu. Rev. Microbiol. 38:339-357.

Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, Kimelman M, Pruitt KD, Resenchuk S, Tatusova T et al. 2012. BioProject and BioSample databases at NCBI: Facilitating capture and organization of metadata. Nucleic Acids Res. 40:D57-63.

Bates JT, Chivian D, Arkin AP. 2011. GLAMM: Genome-linked application for metabolic maps. Nucleic Acids Res. 39:W400-W405.

Bayer TS, Widmaier DM, Temme K, Mirsky EA, Santi DV, Voigt CA. 2009. Synthesis of methyl halides from biomass using engineered microbes. J. Am. Chem. Soc. 131:6508-6515.

Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgård MJ. 2007. Quantitative prediction of cellular metabolism with constraint-based models: The COBRA toolbox. Nat. Protoc. 2:727-738.

Beiko RG. 2011. Telling the whole story in a 10,000-genome world. Biol. Direct 6:34-6150-6-34.

Bendich AJ, Drlica K. 2000. Prokaryotic and eukaryotic chromosomes: What's the difference? Bioessays 22:481-486.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2012. GenBank. Nucleic Acids Res. 41:D36-D42.

Bentley R. 2005. The development of penicillin: Genesis of a famous antibiotic. Perspect. Biol. Med. 48:444-452.

Bentley SD, Parkhill J. 2004. Comparative genomic structure of prokaryotes. Annu. Rev. Genet. 38:771-792.

Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D. 2002. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3 (2). Nature 417:141-147.

Bentley SD, Brown S, Murphy LD, Harris DE, Quail MA, Parkhill J, Barrell BG, McCormick JR, Santamaria RI, Losick R et al. 2004. SCP1, a 356,023 bp linear plasmid adapted to the ecology and developmental biology of its host, *Streptomyces coelicolor* A3(2). Mol. Microbiol. 51:1615-1628.

Berdy J. 1995. Are actinomycetes exhausted as a source of secondary metabolites? Russian Biotechnology 7:3-23.

Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW 3rd, Bulyk ML. 2006. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat. Biotechnol. 24:1429-1435.

Bergmann S, Schümann J, Scherlach K, Lange C, Brakhage AA, Hertweck C. 2007. Genomics-driven discovery of PKS-NRPS hybrid metabolites from *Aspergillus nidulans*. Nat. Chem. Biol. 3:213-217.

Bibb MJ. 2005. Regulation of secondary metabolism in streptomycetes. Curr. Opin. Microbiol. 8:208-215.

Blank LM, Lehmbeck F, Sauer U. 2005. Metabolic-flux and network analysis in fourteen hemiascomycetous yeasts. FEMS Yeast Res. 5:545-558.

Blazeck J, Alper HS. 2012. Promoter engineering: Recent advances in controlling transcription at the most fundamental level. Biotechnol. J. 8:46-58.

Blin K, Medema MH, Kazempour D, Fischbach MA, Breitling R, Takano E, Weber T. 2013. antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res. 41: W204-W212.

Blom J, Albaum SP, Doppmeier D, Puhler A, Vorholter FJ, Zakrzewski M, Goesmann A. 2009. EDGAR: A software framework for the comparative analysis of prokaryotic genomes. BMC Bioinformatics 10:154.

Bode HB, Müller R. 2005. The impact of bacterial genomics on natural product research. Angew. Chem. Int. Ed Engl. 44:6828-6846.

Bode M, Khor S, Ye H, Li MH, Ying JY. 2009. TmPrime: Fast, flexible oligonucleotide design software for gene synthesis. Nucleic Acids Res. 37:W214-21.

Boghigian BA, Pfeifer BA. 2008. Current status, strategies, and potential for the metabolic engineering of heterologous polyketides in *Escherichia coli*. Biotechnol. Lett. 30:1323-1330.

Bond-Watts BB, Bellerose RJ, Chang MC. 2011. Enzyme mechanism as a kinetic control element for designing synthetic biofuel pathways. Nat. Chem. Biol. 7:222-227.

Borodina I, Krabben P, Nielsen J. 2005. Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. Genome Res. 15:820-829.

Brady SF, Simmons L, Kim JH, Schmidt EW. 2009. Metagenomic approaches to natural products from free-living and symbiotic organisms. Nat. Prod. Rep. 26:1488-1503.

Brajtburg J, Powderly WG, Kobayashi GS, Medoff G. 1990. Amphotericin B: Current understanding of mechanisms of action. Antimicrob. Agents Chemother. 34:183-188.

Bray T, Chan P, Bougouffa S, Greaves R, Doig AJ, Warwicker J. 2009. SitesIdentify: A protein functional site prediction tool. BMC Bioinformatics 10:379.

Breitling R, Vitkup D, Barrett MP. 2008. New surveyor tools for charting microbial metabolic maps. Nat. Rev. Microbiol. 6:156-161.

Brenner K, You L, Arnold FH. 2008. Engineering microbial consortia: A new frontier in synthetic biology. Trends Biotechnol. 26:483-489.

Brochado AR, Matos C, Moller BL, Hansen J, Mortensen UH, Patil KR. 2010. Improved vanillin production in baker's yeast through *in silico* design. Microb. Cell. Fact. 9:84.

Brogden RN, Carmine A, Heel RC, Morley PA, Speight TM, Avery GS. 1981. Amoxycillin/clavulanic acid: A review of its antibacterial activity, pharmacokinetics and therapeutic use. Drugs 22:337-362.

Brolle DF, Pape H, Hopwood DA, Kieser T. 1993. Analysis of the transfer region of the *Streptomyces* plasmid SCP2. Mol. Microbiol. 10:157-170.

Brooks JP, Burns WP, Fong SS, Gowen CM, Roberts SB. 2012. Gap detection for genome-scale constraint-based models. Adv. Bioinformatics 2012:323472.

Bumpus SB, Evans BS, Thomas PM, Ntai I, Kelleher NL. 2009. A proteomics approach to discovering natural products and their biosynthetic pathways. Nat. Biotechnol. 27:951-956.

Burgard AP, Pharkya P, Maranas CD. 2003. Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol. Bioeng. 84:647-657.

Caboche S, Pupin M, Leclere V, Fontaine A, Jacques P, Kucherov G. 2008. NORINE: A database of nonribosomal peptides. Nucleic Acids Res. 36:D326-31.

Cai F, Menon BB, Cannon GC, Curry KJ, Shively JM, Heinhorst S. 2009. The pentameric vertex proteins are necessary for the icosahedral carboxysome shell to function as a CO2 leakage barrier. PLoS ONE. 4:e7521.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: Architecture and applications. BMC Bioinformatics 10:421.

Canton B, Labno A, Endy D. 2008. Refinement and standardization of synthetic biological parts and devices. Nat. Biotechnol. 26:787-793.

Carbonell P, Planson AG, Fichera D, Faulon JL. 2011. A retrosynthetic biology approach to metabolic pathway design for therapeutic production. BMC Syst. Biol. 5:122.

Cardinale S, Arkin AP. 2012. Contextualizing context for synthetic biology--identifying causes of failure of synthetic biological systems. Biotechnol. J. 7:856-866.

Carlson R. 2009. The changing economics of DNA synthesis. Nat. Biotechnol. 27:1091-1094.

Challis GL. 2008. Genome mining for novel natural product discovery. J. Med. Chem. 51:2618-2628.

Challis GL, Hopwood DA. 2003. Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by *Streptomyces* species. Proc. Natl. Acad. Sci. U. S. A. 100:14555-14561.

Chan LY, Kosuri S, Endy D. 2005. Refactoring bacteriophage T7. Mol. Syst. Biol. 1:2005.0018.

Chanda A, Roze LV, Kang S, Artymovich KA, Hicks GR, Raikhel NV, Calvo AM, Linz JE. 2009. A key role for vesicles in fungal secondary metabolism. Proc. Natl. Acad. Sci. U. S. A. 106:19533-19538.

Chandran D, Bergmann FT, Sauro HM. 2009. TinkerCell: Modular CAD tool for synthetic biology. J. Biol. Eng. 3:19.

Chang HM, Chen MY, Shieh YT, Bibb MJ, Chen CW. 1996. The *cutRS* signal transduction system of S*treptomyces lividans* represses the biosynthesis of the polyketide antibiotic actinorhodin. Mol. Microbiol. 21:1075-1085.

Chang MC, Eachus RA, Trieu W, Ro DK, Keasling JD. 2007. Engineering *Escherichia coli* for production of functionalized terpenoids using plant P450s. Nat. Chem. Biol. 3:274-277.

Chater KF, Kinashi H. 2007. *Streptomyces* linear plasmids: their discovery, functions, interactions with other replicons, and evolutionary significance. In: Microbial Linear Plasmids, Springer Berlin/Heidelberg, pp. 1-31.

Chechik G, Oh E, Rando O, Weissman J, Regev A, Koller D. 2008. Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network. Nat. Biotechnol. 26:1251-1259.

Chen CW. 2007. *Streptomyces* linear plasmids: replication and telomeres. In: Microbial Linear Plasmids. Berlin/Heidelberg: Springer. pp. 33-61.

Chen CW, Huang CH, Lee HH, Tsai HH, Kirby R. 2002. Once the circle has been broken: Dynamics and evolution of *Streptomyces* chromosomes. Trends Genet. 18:522-529.

Chen CW, Lin YS, Yang YL, Tsou MF, Chang HM, Kieser HM, Hopwood DA. 1994. The linear chromosomes of *Streptomyces*: Structure and dynamics. Actinomycetologica 8:103-112.

Cheng KC, Lin YH, Wu JY, Hwang SCJ. 2003. Enhancement of clavulanic acid production in *Streptomyces clavuligerus* with ornithine feeding. Enzyme Microb. Tech. 32:152-156.

Cho A, Yun H, Park J, Lee S, Park S. 2010. Prediction of novel synthetic pathways for the production of desired chemicals. BMC Syst. Biol. 4:35.

Chopra I, Hawkey PM, Hinton M. 1992. Tetracyclines, molecular and clinical aspects. J. Antimicrob. Chemother. 29:245-277.

Chou C, Chang W, Chiu C, Huang C, Huang H. 2009. FMM: A web server for metabolic pathway reconstruction and comparative analysis. Nucleic Acids Res. 37:W129-W134.

Choulet F, Aigle B, Gallois A, Mangenot S, Gerbaud C, Truong C, Francou FX, Fourrier C, Guerineau M, Decaris B et al. 2006. Evolution of the terminal regions of the *Streptomyces* linear chromosome. Mol. Biol. Evol. 23:2361-2369.

Clardy J, Fischbach MA, Walsh CT. 2006. New antibiotics from bacterial natural products. Nat. Biotechnol. 24:1541-1550.

Conrado RJ, Varner JD, DeLisa MP. 2008. Engineering the spatial organization of metabolic enzymes: Mimicking nature's synergy. Curr. Opin. Biotechnol. 19:492-499.

Conway KR, Boddy CN. 2013. ClusterMine360: A database of microbial PKS/NRPS biosynthesis. Nucleic Acids Res. 41:D402-7.

Cordero OX, Wildschutte H, Kirkup B, Proehl S, Ngo L, Hussain F, Le Roux F, Mincer T, Polz MF. 2012. Ecological populations of bacteria act as socially cohesive units of antibiotic production and resistance. Science 337:1228-1231.

Covert MW, Schilling CH, Famili I, Edwards JS, Goryanin II, Selkov E, Palsson BØ. 2001. Metabolic modeling of microbial strains *in silico*. Trends Biochem. Sci. 26:179-186.

Crawford JM, Clardy J. 2012. Microbial genome mining answers longstanding biosynthetic questions. Proc. Natl. Acad. Sci. U. S. A. 109:7589-7590.

Cvijovic M, Olivares-Hernandez R, Agren R, Dahr N, Vongsangnak W, Nookaew I, Patil KR, Nielsen J. 2010. BioMet toolbox: Genome-wide analysis of metabolism. Nucleic Acids Res. 38:W144-9.

Czar MJ, Cai Y, Peccoud J. 2009. Writing DNA with GenoCAD. Nucleic Acids Res. 37:W40-7.

Dangel V, Westrich L, Smith MC, Heide L, Gust B. 2010. Use of an inducible promoter for antibiotic production in a heterologous host. Appl. Microbiol. Biotechnol. 87:261-269.

Danino T, Mondragon-Palomino O, Tsimring L, Hasty J. 2010. A synchronized quorum of genetic clocks. Nature 463:326-330.

Dano S, Madsen MF, Sorensen PG. 2007. Quantitative characterization of cell synchronization in yeast. Proc. Natl. Acad. Sci. U. S. A. 104:12732-12736.

Darling AC, Mau B, Blattner FR, Perna NT. 2004. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14:1394-1403.

Davies J, Ryan KS. 2012. Introducing the parvome: Bioactive compounds in the microbial world. ACS Chem. Biol. 7:252-259.

Davis JH, Rubin AJ, Sauer RT. 2011. Design, construction and characterization of a set of insulated bacterial promoters. Nucleic Acids Res. 39:1131-1141.

de Jong A, van Heel AJ, Kok J, Kuipers OP. 2010. BAGEL2: Mining for bacteriocins in genomic data. Nucleic Acids Res. 38:W647-51.

DeJongh M, Formsma K, Boillot P, Gould J, Rycenga M, Best A. 2007. Toward the automated generation of genome-scale metabolic networks in the SEED. BMC Bioinformatics 8:139.

Delcher AL, Bratke KA, Powers EC, Salzberg SL. 2007. Identifying bacterial genes and endosymbiont DNA with glimmer. Bioinformatics 23:673-679.

Delves-Broughton J. 1990. Nisin and its uses as a food preservative. Food Technol. 44:100-117.

Demange P, Bateman A, Dell A, Abdallah MA. 1988. Structure of azotobactin D, a siderophore of *Azotobacter vinelandii* strain D (CCM 289). Biochemistry (N. Y.) 27:2745-2752.

Despalins A, Marsit S, Oberto J. 2011. Absynte: A web tool to analyze the evolution of orthologous archaeal and bacterial gene clusters. Bioinformatics 27:2905-2906.

Dhote V, Gupta S, Reynolds KA. 2008. An O-phosphotransferase catalyzes phosphorylation of hygromycin A in the antibiotic-producing organism S*treptomyces hygroscopicus*. Antimicrob. Agents Chemother. 52:3580-3588.

Dietrich JA, Yoshikuni Y, Fisher KJ, Woolard FX, Ockey D, McPhee DJ, Renninger NS, Chang MC, Baker D, Keasling JD. 2009. A novel semi-biosynthetic route for artemisinin production using engineered substrate-promiscuous P450(BM3). ACS Chem. Biol. 4:261-267.

Dixon N, Duncan JN, Geerlings T, Dunstan MS, McCarthy JE, Leys D, Micklefield J. 2010. Reengineering orthogonally selective riboswitches. Proc. Natl. Acad. Sci. U. S. A. 107:2830-2835.

Dobrindt U, Hochhut B, Hentschel U, Hacker J. 2004. Genomic islands in pathogenic and environmental microorganisms. Nat. Rev. Microbiol. 2:414-424.

Donadio S, Monciardini P, Sosio M. 2007. Polyketide synthases and nonribosomal peptide synthetases: The emerging view from bacterial genomics. Nat. Prod. Rep. 24:1073-1109.

Donadio S, Sosio M, Stegmann E, Weber T, Wohlleben W. 2005. Comparative analysis and insights into the evolution of gene clusters for glycopeptide antibiotic biosynthesis. Mol. Genet. Genomics 274:40-50.

Drepper T, Gross S, Yakunin AF, Hallenbeck PC, Masepohl B, Klipp W. 2003. Role of GlnB and GlnK in ammonium control of both nitrogenase systems in the phototrophic bacterium *Rhodobacter capsulatus*. Microbiology 149:2203-2212.

Du L, Villarreal S, Forster AC. 2012. Multigene expression *in vivo*: Supremacy of large versus small terminators for T7 RNA polymerase. Biotechnol. Bioeng. 109:1043-1050.

Du L, Gao R, Forster AC. 2009. Engineering multigene expression *in vitro* and *in vivo* with small terminators for T7 RNA polymerase. Biotechnol. Bioeng. 104:1189-1196.

Duarte NC, Herrgård MJ, Palsson BØ. 2004. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. Genome Res. 14:1298-1309.

Dueber JE, Wu GC, Malmirchegini GR, Moon TS, Petzold CJ, Ullal AV, Prather KL, Keasling JD. 2009. Synthetic protein scaffolds provide modular control over metabolic flux. Nat. Biotechnol. 27:753-759.

Dunlop MJ, Dossani ZY, Szmidt HL, Chu HC, Lee TS, Keasling JD, Hadi MZ, Mukhopadhyay A. 2011. Engineering microbial biofuel tolerance and export using efflux pumps. Mol. Syst. Biol. 7:487.

Durot M, Bourguignon P-, Schachter V. 2009. Genome-scale models of bacterial metabolism: Reconstruction and applications. FEMS Microbiol. Rev. 33:164-190.

Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput. Biol. 7:e1002195.

Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. Genome Inform. 23:205-211.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792-1797.

Edwards JS, Ibarra RU, Palsson BØ. 2001. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. Nat. Biotechnol. 19:125-130.

Edwards JS, Ramakrishna R, Schilling CH. 1999. Metabolic flux balance analysis. Metab. Eng. :13-57.

Ellis T, Adie T, Baldwin GS. 2011. DNA assembly for synthetic biology: From parts to pathways and beyond. Integr. Biol. (Camb) 3:109-118.

Ellis T, Wang X, Collins JJ. 2009. Diversity-based, model-guided construction of synthetic gene networks with predicted functions. Nat. Biotechnol. 27:465-471.

Elowitz MB, Surette MG, Wolf PE, Stock JB, Leibler S. 1999. Protein mobility in the cytoplasm of *Escherichia coli*. J. Bacteriol. 181:197-203.

English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC et al. 2012. Mind the gap: Upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS ONE 7:e47768.

Erb A, Weiss H, Harle J, Bechthold A. 2009. A bacterial glycosyltransferase gene toolbox: Generation and applications. Phytochemistry 70:1812-1821.

Evans RH,Jr., Ax H, Jacoby A, Williams TH, Jenkins E, Scannell JP. 1983. Ro 22-5417, a new clavam antibiotic from *Streptomyces clavuligerus*. II. fermentation, isolation and structure. J. Antibiot. (Tokyo) 36:213-216.

Evers ME, Trip H, van den Berg MA, Bovenberg RA, Driessen AJ. 2004. Compartmentalization and transport in beta-lactam antibiotics biosynthesis. Adv. Biochem. Eng. Biotechnol. 88:111-135.

Fan C, Cheng S, Liu Y, Escobar CM, Crowley CS, Jefferson RE, Yeates TO, Bobik TA. 2010. Short N-terminal sequences package proteins into bacterial microcompartments. Proc. Natl. Acad. Sci. U. S. A. 107:7509-7514.

Farmer WR, Liao JC. 2000. Improving lycopene production in *Escherichia coli* by engineering metabolic control. Nat. Biotechnol. 18:533-537.

Federhen S. 2012. The NCBI taxonomy database. Nucleic Acids Res. 40:D136-43.

Fedorova ND, Moktali V, Medema MH. 2012. Bioinformatics approaches and software for detection of secondary metabolic gene clusters. Methods Mol. Biol. 944:23-45.

Feist AM, Palsson BØ. 2010. The biomass objective function. Curr. Opin. Microbiol. 13:344-349.

Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ. 2009. Reconstruction of biochemical networks in microorganisms. Nat. Rev. Microbiol. 7:129-143.

Feist AM, Zielinski DC, Orth JD, Schellenberger J, Herrgård MJ, Palsson BØ. 2010. Model-driven evaluation of the production potential for growth-coupled products of *Escherichia coli*. Metab. Eng. 12:173-186.

Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ. 2007. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol. Syst. Biol. 3:121.

Field B, Osbourn AE. 2008. Metabolic diversification--independent assembly of operon-like gene clusters in different plants. Science 320:543-547.

Field D, Amaral-Zettler L, Cochrane G, Cole JR, Dawyndt P, Garrity GM, Gilbert J, Glockner FO, Hirschman L, Karsch-Mizrachi I et al. 2011. The Genomic Standards Consortium. PLoS Biol. 9:e1001088.

Finley S, Broadbelt L, Hatzimanikatis V. 2010. *In silico* feasibility of novel biodegradation pathways for 1,2,4-trichlorobenzene. BMC Syst. Biol. 4:7.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K et al. 2010. The Pfam protein families database. Nucleic Acids Res. 38:D211-22.

Fischbach M, Voigt CA. 2010. Prokaryotic gene clusters: A rich toolbox for synthetic biology. Biotechnol. J. 5:1277-1296.

Fischbach MA, Walsh CT. 2009. Antibiotics for emerging pathogens. Science 325:1089-1093.

Fischbach MA, Walsh CT. 2006. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: Logic, machinery, and mechanisms. Chem. Rev. 106:3468-3496.

Fischbach MA, Walsh CT, Clardy J. 2008. The evolution of gene collectives: How natural selection drives chemical innovation. Proc. Natl. Acad. Sci. U. S. A. 105:4601-4608.

Fleischmann R, Adams M, White O, Clayton R, Kirkness E, Kerlavage A, Bult C, Tomb J, Dougherty B, Merrick J et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* rd. Science 269:496-512.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S et al. 2012. Ensembl 2012. Nucleic Acids Res. 40:D84-90.

Fong C, Rohmer L, Radey M, Wasnick M, Brittnacher MJ. 2008. PSAT: A web tool to compare genomic neighborhoods of multiple prokaryotic genomes. BMC Bioinformatics 9:170.

Fordyce PM, Gerber D, Tran D, Zheng J, Li H, DeRisi JL, Quake SR. 2010. *De novo* identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. Nat. Biotechnol. 28:970-975.

Fowler M, Beck K. 1999. Refactoring: Improving the design of existing code. Addison-Wesley Professional.

Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C et al. 2013. STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 41:D808-15.

Francke C, Kerkhoven R, Wels M, Siezen RJ. 2008. A generic approach to identify transcription factor-specific operator motifs; inferences for LacI-family mediated regulation in *Lactobacillus plantarum* WCFS1. BMC Genomics 9:145.

Franke J, Ishida K, Hertweck C. 2012. Genomics-driven discovery of burkholderic acid, a noncanonical, cryptic polyketide from human pathogenic *Burkholderia* species. Angew. Chem. Int. Ed Engl. 51:11611-11615.

Freeman MF, Gurgui C, Helf MJ, Morinaka BI, Uria AR, Oldham NJ, Sahl HG, Matsunaga S, Piel J. 2012. Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. Science 338:387-390.

Freitag A, Mendez C, Salas JA, Kammerer B, Li SM, Heide L. 2006. Metabolic engineering of the heterologous production of clorobiocin derivatives and elloramycin in *Streptomyces coelicolor* M512. Metab. Eng. 8:653-661.

Frost LS, Leplae R, Summers AO, Toussaint A. 2005. Mobile genetic elements: The agents of open source evolution. Nat. Rev. Microbiol. 3:722-732.

Fung E, Wong WW, Suen JK, Bulter T, Lee SG, Liao JC. 2005. A synthetic gene-metabolic oscillator. Nature 435:118-122.

Galdzicki M, Rodriguez C, Chandran D, Sauro HM, Gennari JH. 2011. Standard biological parts knowledgebase. PLoS ONE 6:e17005.

Galvao TC, de L,V. 2006. Transcriptional regulators a la carte: Engineering new effector specificities in bacterial regulatory proteins. Curr. Opin. Biotechnol. 17:34-42.

Garcia-Fernandez J. 2005. The genesis and evolution of homeobox gene clusters. Nat. Rev. Genet. 6:881-892.

Gerber NN, Lechevalier HA. 1965. Geosmin, an earthly-smelling substance isolated from actinomycetes. Appl. Microbiol. 13:935-938.

Gerke J, Bayram O, Feussner K, Landesfeind M, Shelest E, Feussner I, Braus GH. 2012. Breaking the silence: Protein stabilization uncovers silenced biosynthetic gene clusters in the fungus *Aspergillus nidulans*. Appl. Environ. Microbiol. 78:8234-8244.

Gevorgyan A, Bushell ME, Avignone-Rossa C, Kierzek AM. 2011. SurreyFBA: A command line tool and graphics user interface for constraint-based modeling of genome-scale metabolic reaction networks. Bioinformatics 27:433-434.

Ghorbel S, Kormanec J, Artus A, Virolle MJ. 2006. Transcriptional studies and regulatory interactions between the *phoR-phoP* operon and the *phoU*, *mtpA*, and *ppk* genes of S*treptomyces lividans* TK24. J. Bacteriol. 188:677-686.

Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA,III, Smith HO. 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods 6:343-345.

Gibson DG, Benders GA, ndrews-Pfannkoch C, Denisova EA, Baden-Tillson H, Zaveri J, Stockwell TB, Brownley A, Thomas DW, Algire MA et al. 2008. Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. Science 319:1215-1220.

Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, Algire MA, Benders GA, Montague MG, Ma L, Moodie MM et al. 2010. Creation of a bacterial cell controlled by a chemically synthesized genome. Science 329:52-56.

Giessen TW, Marahiel MA. 2012. Ribosome-independent biosynthesis of biologically active peptides: Application of synthetic biology to generate structural diversity. FEBS Lett. 586:2065-2075.

Gomez-Escribano JP, Bibb MJ. 2011. Engineering *Streptomyces coelicolor* for heterologous expression of secondary metabolite gene clusters. Microb. Biotechnol. 4:207-215.

Gonzalez-Lergier J, Broadbelt LJ, Hatzimanikatis V. 2005. Theoretical considerations and computational analysis of the complexity in polyketide synthesis pathways. J. Am. Chem. Soc. 127:9930-9938.

Gottelt M, Kol S, Gomez-Escribano JP, Bibb M, Takano E. 2010. Deletion of a regulatory gene within the *cpk* gene cluster reveals novel antibacterial activity in *Streptomyces coelicolor* A3(2). Microbiology 156:2343-2353.

Goyal K, Mohanty D, Mande SC. 2007. PAR-3D: A server to predict protein active site residues. Nucleic Acids Res. 35:W503-5.

Graham RM. 1994. Cyclosporine: Mechanisms of action and toxicity. Cleve. Clin. J. Med. 61:308-313.

Gravius B, Glocker D, Pigac J, Pandza K, Hranueli D, Cullum J. 1994. The 387 kb linear plasmid pPZG101 of *Streptomyces rimosus* and its interactions with the chromosome. Microbiology 140:2271-2277.

Grigoriev A. 1998. Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res. 26:2286-2290.

Gullo VP, McAlpine J, Lam KS, Baker D, Petersen F. 2006. Drug discovery from natural products. J. Ind. Microbiol. Biotechnol. 33:523-531.

Guthals A, Watrous JD, Dorrestein PC, Bandeira N. 2012. The spectral networks paradigm in high throughput mass spectrometry. Mol. Biosyst 8:2535-2544.

Hall BG. 2006. Simple and accurate estimation of ancestral protein sequences. Proc. Natl. Acad. Sci. U. S. A. 103:5431-5436.

Hanai T, Atsumi S, Liao JC. 2007. Engineered synthetic pathway for isopropanol production in *Escherichia coli*. Appl. Environ. Microbiol. 73:7814-7818.

Hashizume H, Igarashi M, Hattori S, Hori M, Hamada M, Takeuchi T. 2001. Tripropeptins, novel antimicrobial agents produced by *Lysobacter* sp. I. taxonomy, isolation and biological activities. J. Antibiot. (Tokyo) 54:1054-1059.

Hassan KA, Johnson A, Shaffer BT, Ren Q, Kidarsa TA, Elbourne LD, Hartney S, Duboy R, Goebel NC, Zabriskie TM et al. 2010. Inactivation of the GacA response regulator in *Pseudomonas fluorescens* Pf-5 has far-reaching transcriptomic consequences. Environ. Microbiol. 12:899-915.

Hatzimanikatis V, Li C, Ionita JA, Henry CS, Jankowski MD, Broadbelt LJ. 2005. Exploring the diversity of complex metabolic networks. Bioinformatics 21:1603-1609.

Heath AP, Bennett GN, Kavraki LE. 2010. Finding metabolic pathways using atom tracking. Bioinformatics 26:1548-1555.

Heneghan MN, Yakasai AA, Halo LM, Song Z, Bailey AM, Simpson TJ, Cox RJ, Lazarus CM. 2010. First heterologous reconstruction of a complete functional fungal biosynthetic multigene cluster. Chembiochem 11:1508-1512.

Henry CS, Broadbelt LJ, Hatzimanikatis V. 2010. Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate. Biotechnol. Bioeng. 106:462-473.

Henry CS, Broadbelt LJ, Hatzimanikatis V. 2007. Thermodynamics-based metabolic flux analysis. Biophys. J. 92:1792-1805.

Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. 2010. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat. Biotechnol. 28:977-982.

Higgens CE, Kastner RE. 1971. *Streptomyces clavuligerus* sp. nov., a ß-lactam antibiotic producer. Int. J. Syst. Bacteriol. 21:326-331.

Hinnebusch J, Tilly K. 1993. Linear plasmids and chromosomes in bacteria. Mol. Microbiol. 10:917-922.

Holtz WJ, Keasling JD. 2010. Engineering static and dynamic control of synthetic pathways. Cell 140:19-23.

Hoover DM, Lubkowski J. 2002. DNAWorks: An automated method for designing oligonucleotides for PCR-based gene synthesis. Nucleic Acids Res. 30:e43.

Hopwood DA. 2007. *Streptomyces* in nature and medicine: The antibiotic makers. New York: Oxford University Press.

Hopwood DA. 2006. Soil to genomics: The *Streptomyces* chromosome. Annu. Rev. Genet. 40:1-23.

Hou BK, Wackett LP, Ellis LBM. 2003. Microbial pathway prediction: A functional group approach. J. Chem. Inf. Comput. Sci. 43:1051-1057.

Hsu CC, Chen CW. 2010. Linear plasmid SLP2 is maintained by partitioning, intrahyphal spread, and conjugal transfer in *Streptomyces*. J. Bacteriol. 192:307-315.

Huang CH, Tsai HH, Tsay YG, Chien YN, Wang SL, Cheng MY, Ke CH, Chen CW. 2007. The telomere system of the *Streptomyces* linear plasmid SCP1 represents a novel class. Mol. Microbiol. 63:1710-1718.

Hung TV, Malla S, Park BC, Liou K, Lee HC, Sohng JK. 2007. Enhancement of clavulanic acid by replicative and integrative expression of *ccaR* and *cas2* in *Streptomyces clavuligerus* NRRL3585. J. Microbiol. Biotechnol. 17:1538-1545.

Ibrahim A, Yang L, Johnston C, Liu X, Ma B, Magarvey NA. 2012. Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. Proc. Natl. Acad. Sci. U. S. A. 109:19196-19201.

Ichikawa N, Oguchi A, Ikeda H, Ishikawa J, Kitani S, Watanabe Y, Nakamura S, Katano Y, Kishi E, Sasagawa M et al. 2010. Genome sequence of *Kitasatospora setae* NBRC 14216T: An evolutionary snapshot of the family *Streptomycetaceae*. DNA Res. 17:393-406.

Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, Sakaki Y, Hattori M, Omura S. 2003. Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. Nat. Biotechnol. 21:526-531.

Irmscher G, Bovermann G, Boheim G, Jung G. 1978. Trichotoxin A-40, a new membrane-exciting peptide. part A. isolation, characterization and conformation. Biochim. Biophys. Acta 507:470-484.

Ishizuka H, Horinouchi S, Kieser HM, Hopwood DA, Beppu T. 1992. A putative two-component regulatory system involved in secondary metabolism in *Streptomyces* spp. J. Bacteriol. 174:7585-7594.

Itaya M, Kaneko S. 2010. Integration of stable extracellular DNA released from *Escherichia coli* into the *Bacillus subtilis* genome vector by culture mix method. Nucleic Acids Res. 38:2551-2557.

Itaya M, Fujita K, Kuroki A, Tsuge K. 2008. Bottom-up genome assembly using the *Bacillus subtilis* genome vector. Nat. Methods 5:41-43.

Ives PR, Bushell ME. 1997. Manipulation of the physiology of clavulanic acid production in *Streptomyces clavuligerus*. Microbiology 143:3573-3579.

Jankowski MD, Henry CS, Broadbelt LJ, Hatzimanikatis V. 2008. Group contribution method for thermodynamic analysis of complex metabolic networks. Biophys. J. 95:1487-1499.

Jenke-Kodama H, Borner T, Dittmann E. 2006. Natural biocombinatorics in the polyketide synthase genes of the actinobacterium *Streptomyces avermitilis*. PLoS. Comput. Biol. 2:e132.

Jensen PR, Williams PG, Oh DC, Zeigler L, Fenical W. 2007. Species-specific secondary metabolite production in marine actinomycetes of the genus *Salinispora*. Appl. Environ. Microbiol. 73:1146-1152.

Jnawali HN, Lee HC, Sohng JK. 2010. Enhancement of clavulanic acid production by expressing regulatory genes in *gap* gene deletion mutant of *Streptomyces clavuligerus* NRRL3585. J. Microbiol. Biotechnol. 20:146-152.

Kanehisa M, Goto S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28:27-30.

Kanehisa M, Goto S, Kawashima S, Nakaya A. 2002. The KEGG databases at GenomeNet. Nucleic Acids Res. 30:42-46.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. 2004. The KEGG resource for deciphering the genome. Nucleic Acids Res. 32:D277-80.

Kealey JT, Liu L, Santi DV, Betlach MC, Barr PJ. 1998. Production of a polyketide natural product in nonpolyketide-producing prokaryotic and eukaryotic hosts. Proc. Natl. Acad. Sci. U. S. A. 95:505-509.

Keasling JD. 2008. Synthetic biology for synthetic chemistry. ACS Chem. Biol. 3:64-76.

Kenig M, Reading C. 1979. Holomycin and an antibiotic (MM 19290) related to tunicamycin, metabolites of *Streptomyces clavuligerus*. J. Antibiot. (Tokyo) 32:549-554.

Kersten RD, Yang YL, Xu Y, Cimermancic P, Nam SJ, Fenical W, Fischbach MA, Moore BS, Dorrestein PC. 2011. A mass spectrometry-guided genome mining approach for natural product peptidogenomics. Nat. Chem. Biol. 7:794-802.

Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, Fedorova ND. 2010. SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genet. Biol. 47:736-741.

Khalil AS, Collins JJ. 2010. Synthetic biology: Applications come of age. Nat. Rev. Genet. 11:367-379.

Khosla C, Kapur S, Cane DE. 2009. Revisiting the modularity of modular polyketide synthases. Curr. Opin. Chem. Biol. 13:135-143.

Kim HJ, Boedicker JQ, Choi JW, Ismagilov RF. 2008. Defined spatial structure stabilizes a synthetic multispecies bacterial community. Proc. Natl. Acad. Sci. U. S. A. 105:18188-18193.

Kim HS, Park YI. 2008. Isolation and identification of a novel microorganism producing the immunosuppressant tacrolimus. J. Biosci. Bioeng. 105:418-421.

Kim HS, Lee YJ, Lee CK, Choi SU, Yeo SH, Hwang YI, Yu TS, Kinoshita H, Nihira T. 2004. Cloning and characterization of a gene encoding the γ-butyrolactone autoregulator receptor from *Streptomyces clavuligerus*. Arch. Microbiol. 182:44-50.

Kim TY, Sohn SB, Kim YB, Kim WJ, Lee SY. 2012. Recent advances in reconstruction and applications of genome-scale metabolic models. Curr. Opin. Biotechnol. 23:617-623.

Kinashi H. 2008. Antibiotic production, linear plasmids and linear chromosomes in *Streptomyces*. Actinomycetologica 22:20-29.

Kinashi H, Shimaji-Murayama M, Hanafusa T. 1992. Integration of SCP1, a giant linear plasmid, into the *Streptomyces coelicolor* chromosome. Gene 115:35-41.

Kirby R. 1978. An unstable genetic element affecting the production of the antibiotic holomycin by *Streptomyces clavuligerus*. FEMS Microbiol. Lett. 3:283-286.

Kirby R, Hopwood DA. 1977. Genetic determination of methylenomycin synthesis by the SCP1 plasmid of *Streptomyces coelicolor* A3(2). J. Gen. Microbiol. 98:239-252.

Klassen JL, Currie CR. 2012. Gene fragmentation in bacterial draft genomes: Extent, consequences and mitigation. BMC Genomics 13:14-2164-13-14.

Kodumal SJ, Patel KG, Reid R, Menzella HG, Welch M, Santi DV. 2004. Total synthesis of long DNA sequences: Synthesis of a contiguous 32-kb polyketide synthase gene cluster. Proc. Natl. Acad. Sci. U. S. A. 101:15573-15578.

Komatsu M, Uchiyama T, Omura S, Cane DE, Ikeda H. 2010. Genome-minimized *Streptomyces* host for the heterologous expression of secondary metabolism. Proc. Natl. Acad. Sci. U. S. A. 107:2646-2651.

Koski LB, Gray MW, Lang BF, Burger G. 2005. AutoFACT: An automatic functional annotation and classification tool. BMC Bioinformatics 6:151.

Kosuri S, Eroshenko N, Leproust EM, Super M, Way J, Li JB, Church GM. 2010. Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. Nat. Biotechnol. 28:1295-1299.

Krumsiek J, Arnold R, Rattei T. 2007. Gepard: A rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23:1026-1028.

Kwok R. 2010. Five hard truths for synthetic biology. Nature 463:288-290.

Lartigue C, Vashee S, Algire MA, Chuang RY, Benders GA, Ma L, Noskov VN, Denisova EA, Gibson DG, ssad-Garcia N et al. 2009. Creating bacterial strains from genomes that have been cloned and engineered in yeast. Science 325:1693-1696.

Lasken RS. 2012. Genomic sequencing of uncultured microorganisms from single cells. Nat. Rev. Microbiol. 10:631-640.

Latendresse M, Krummenacker M, Trupp M, Karp PD. 2012. Construction and completion of flux balance models from pathway databases. Bioinformatics 28:388-396.

Latreille P, Norton S, Goldman BS, Henkhaus J, Miller N, Barbazuk B, Bode HB, Darby C, Du Z, Forst S et al. 2007. Optical mapping as a routine tool for bacterial genome sequence finishing. BMC Genomics 8:321.

Laureti L, Song L, Huang S, Corre C, Leblond P, Challis GL, Aigle B. 2011. Identification of a bioactive 51-membered macrolide complex by activation of a silent polyketide synthase in *Streptomyces ambofaciens*. Proc. Natl. Acad. Sci. U. S. A. 108:6258-6263.

Lautru S, Deeth RJ, Bailey LM, Challis GL. 2005. Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. Nat. Chem. Biol. 1:265-269.

Le Fevre F, Smidtas S, Combe C, Durot M, d'Alche-Buc F, Schachter V. 2009. CycSim--an online tool for exploring and experimenting with genome-scale metabolic models. Bioinformatics 25:1987-1988.

Lee FC, Pandu Rangaiah G, Lee DY. 2010. Modeling and optimization of a multi-product biosynthesis factory for multiple objectives. Metab. Eng. 12:251-267.

Lee PA, Dymond JS, Scheifele LZ, Richardson SM, Foelber KJ, Boeke JD, Bader JS. 2010. CLONEQC: Lightweight sequence verification for synthetic biology. Nucleic Acids Res. 38:2617-2623.

Lehmann J, Stadler PF, Prohaska SJ. 2008. SynBlast: Assisting the analysis of conserved synteny information. BMC Bioinformatics 9:351.

Lentzen G, Schwarz T. 2006. Extremolytes: Natural compounds from extremophiles for versatile applications. Appl. Microbiol. Biotechnol. 72:623-634.

Letunic I, Bork P. 2011. Interactive tree of life v2: Online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 39:W475-8.

Letunic I, Doerks T, Bork P. 2009. SMART 6: Recent updates and new developments. Nucleic Acids Res. 37:D229-32.

Letunic I, Yamada T, Kanehisa M, Bork P. 2008. iPath: Interactive exploration of biochemical pathways and networks. Trends Biochem. Sci. 33:101-103.

Letzel AC, Pidot SJ, Hertweck C. 2013. A genomic approach to the cryptic secondary metabolome of the anaerobic world. Nat. Prod. Rep. 30:392-428.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G et al. 2007. The diploid genome sequence of an individual human. PLoS. Biol. 5:e254.

Lewis NE, Nagarajan H, Palsson BØ. 2012. Constraining the metabolic genotype-phenotype relationship using a phylogeny of *in silico* methods. Nat. Rev. Microbiol. 10:291-305.

Li JW, Vederas JC. 2009. Drug discovery and natural products: End of an era or an endless frontier? Science 325:161-165.

Li L, Stoeckert CJ,Jr., Roos DS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178-2189.

Li MH, Ung PM, Zajkowski J, Garneau-Tsodikova S, Sherman DH. 2009. Automated genome mining for natural products. BMC Bioinformatics 10:185.

Li R, Townsend CA. 2006. Rational strain improvement for enhanced clavulanic acid production by genetic engineering of the glycolytic pathway in *Streptomyces clavuligerus*. Metab. Eng. 8:240-252.

Liao D. 1999. Concerted evolution: Molecular mechanism and biological implications. Am. J. Hum. Genet. 64:24-30.

Liao G, Li J, Li L, Yang H, Tian Y, Tan H. 2010. Cloning, reassembling and integration of the entire nikkomycin biosynthetic gene cluster into *Streptomyces ansochromogenes* lead to an improved nikkomycin production. Microb. Cell. Fact. 9:6.

Lim HN, Lee Y, Hussein R. 2011. Fundamental relationship between operon organization and gene expression. Proc. Natl. Acad. Sci. U. S. A. 108:10626-10631.

Lin K, Zhu L, Zhang DY. 2006. An initial strategy for comparing proteins at the domain architecture level. Bioinformatics 22:2081-2086.

Liras P, Gomez-Escribano JP, Santamarta I. 2008. Regulatory mechanisms controlling antibiotic production in *Streptomyces clavuligerus*. J. Ind. Microbiol. Biotechnol. 35:667-676.

Liu G, Chater KF, Chandra G, Niu G, Tan H. 2013. Molecular regulation of antibiotic biosynthesis in *Streptomyces*. Microbiol. Mol. Biol. Rev. 77:112-143.

Liu W, Christenson SD, Standage S, Shen B. 2002. Biosynthesis of the enediyne antitumor antibiotic C-1027. Science 297:1170-1173.

Liu WT, Yang YL, Xu Y, Lamsa A, Haste NM, Yang JY, Ng J, Gonzalez D, Ellermeier CD, Straight PD et al. 2010. Imaging mass spectrometry of intraspecies metabolic exchange revealed the cannibalistic factors of *Bacillus subtilis*. Proc. Natl. Acad. Sci. U. S. A. 107:16286-16290.

Lopez-Garcia MT, Santamarta I, Liras P. 2010. Morfological differentiation and clavulanic acid formation are affected in an *S. clavuligerus* Δ*adpA*-deleted mutant. Microbiology 156:2354-2365.

Lou C, Stanton B, Chen YJ, Munsky B, Voigt CA. 2012. Ribozyme-based insulator parts buffer synthetic circuits from genetic context. Nat. Biotechnol. 30:1137-1142.

Lu TK, Khalil AS, Collins JJ. 2009. Next-generation synthetic gene networks. Nat. Biotechnol. 27:1139-1150.

Lu Y, Wang W, Shu D, Zhang W, Chen L, Qin Z, Yang S, Jiang W. 2007. Characterization of a novel two-component regulatory system involved in the regulation of both actinorhodin and a type I polyketide in *Streptomyces coelicolor*. Appl. Microbiol. Biotechnol. 77:625-635.

Luzhetskyy A, Mendez C, Salas JA, Bechthold A. 2008. Glycosyltransferases, important tools for drug design. Curr. Top. Med. Chem. 8:680-709.

Ma H, Zeng AP. 2003. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics 19:270-277.

Ma S, Tang N, Tian J. 2012. DNA synthesis, assembly and applications in synthetic biology. Curr. Opin. Chem. Biol. 16:260-267.

Machida M, Asai K, Sano M, Tanaka T, Kumagai T, Terai G, Kusumoto K, Arima T, Akita O, Kashiwagi Y. 2005. Genome sequencing and analysis of *Aspergillus oryzae*. Nature 438:1157-1161.

Majoros WH, Pertea M, Salzberg SL. 2004. TigrScan and GlimmerHMM: Two open source *ab initio* eukaryotic gene-finders. Bioinformatics 20:2878-2879.

Mallika V, Sivakumar KC, Jaichand S, Soniya EV. 2010. Kernel based machine learning algorithm for the efficient prediction of type III polyketide synthase family of proteins. J. Integr. Bioinform 7:143.

Malmberg LH, Hu WS. 1992. Identification of rate-limiting steps in cephalosporin C biosynthesis in *Cephalosporium acremonium*: A theoretical analysis. Appl. Microbiol. Biotechnol. 38:122-128.

Marchisio MA, Stelling J. 2009. Computational design tools for synthetic biology. Curr. Opin. Biotechnol. 20:479-485.

Marinelli F. 2009. From microbial products to novel drugs that target a multitude of disease indications. Methods Enzymol. 458:29-58.

Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P et al. 2012. IMG: The integrated microbial genomes database and comparative analysis system. Nucleic Acids Res. 40:D115-22.

Marler,R.T., Arora, J.S. 2004. Survey of multi-objective optimization methods for engineering. Struct. Multidiscip. O. 26:369-395.

Martin CH, Nielsen DR, Solomon KV, Prather KL. 2009. Synthetic metabolism: Engineering biology at the protein and pathway scales. Chem. Biol. 16:277-286.

Martin FJ, McInerney JO. 2009. Recurring cluster and operon assembly for phenylacetate degradation genes. BMC Evol. Biol. 9:36-2148-9-36.

Martin JF, Liras P. 2010. Engineering of regulatory cascades and networks controlling antibiotic biosynthesis in *Streptomyces*. Curr. Opin. Microbiol. 13:263-273.

Martin JF, Liras P. 1989. Enzymes involved in penicillin, cephalosporin and cephamycin biosynthesis. Adv. Biochem. Eng. Biotechnol. 39:153-187.

Martin VJJ, Pitera DJ, Withers ST, Newman JD, Keasling JD. 2003. Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. Nat Biotech 21:796-802.

Martinez JS, Carter-Franklin JN, Mann EL, Martin JD, Haygood MG, Butler A. 2003. Structure and membrane affinity of a suite of amphiphilic siderophores produced by a marine bacterium. Proc. Natl. Acad. Sci. U. S. A. 100:3754-3759.

Matzas M, Stahler PF, Kefer N, Siebelt N, Boisguerin V, Leonard JT, Keller A, Stahler CF, Haberle P, Gharizadeh B et al. 2010. High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing. Nat. Biotechnol. 28:1291-1294.

Mavromatis K, Chu K, Ivanova N, Hooper SD, Markowitz VM, Kyrpides NC. 2009. Gene context analysis in the integrated microbial genomes (IMG) data management system. PLoS ONE 4:e7979.

Mavrovouniotis M, Stephanopoulos G, Stephanopoulos G. 1992. Synthesis of biochemical production routes. Comput. Chem. Eng. 16:605-619.

McDermott JC, Ben-Aziz A, Singh RK, Britton G, Goodwin TW. 1973. Recent studies of carotenoid biosynthesis in bacteria. Pure Appl. Chem. 35:29-45.

McLeod MP, Warren RL, Hsiao WW, Araki N, Myhre M, Fernandes C, Miyazawa D, Wong W, Lillquist AL, Wang D et al. 2006. The complete genome of *Rhodococcus* sp. RHA1 provides insights into a catabolic powerhouse. Proc. Natl. Acad. Sci. U. S. A. 103:15582-15587.

McShan DC, Rao S, Shah I. 2003. PathMiner: Predicting metabolic pathways by heuristic search. Bioinformatics 19:1692-1698.

Medema MH, Takano E, Breitling R. 2013. Detecting sequence homology at the gene cluster level with MultiGeneBlast. Mol. Biol. Evol. 30:1218-1223.

Medema MH, van Raaphorst R, Takano E, Breitling R. 2012. Computational tools for the synthetic design of biochemical pathways. Nat. Rev. Microbiol. 10:191-202.

Medema MH, Breitling R, Bovenberg R, Takano E. 2011a. Exploiting plug-and-play synthetic biology for drug discovery and production in microorganisms. Nat. Rev. Microbiol. 9:131-137.

Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R. 2011b. antiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 39:W339-46.

Medema MH, Alam MT, Heijne WH, van den Berg MA, Müller U, Trefzer A, Bovenberg RA, Breitling R, Takano E. 2011c. Genome-wide gene expression changes in an industrial clavulanic acid overproduction strain of *Streptomyces clavuligerus*. Microb. Biotechnol. 4:300-305.

Medema MH, Trefzer A, Kovalchuk A, van den Berg M, Müller U, Heijne W, Wu L, Alam MT, Ronning CM, Nierman WC et al. 2010. The sequence of a 1.8-mb bacterial linear plasmid reveals a rich evolutionary reservoir of secondary metabolic pathways. Genome Biol. Evol. 2:212-224.

Menzella HG, Reeves CD. 2007. Combinatorial biosynthesis for drug development. Curr. Opin. Microbiol. 10:238-245.

Menzella HG, Carney JR, Santi DV. 2007. Rational design and assembly of synthetic trimodular polyketide synthases. Chem. Biol. 14:143-151.

Menzella HG, Reisinger SJ, Welch M, Kealey JT, Kennedy J, Reid R, Tran CQ, Santi DV. 2006. Redesign, synthesis and functional expression of the 6-deoxyerythronolide B polyketide synthase gene cluster. J. Ind. Microbiol. Biotechnol. 33:22-28.

Menzella HG, Reid R, Carney JR, Chandran SS, Reisinger SJ, Patel KG, Hopwood DA, Santi DV. 2005. Combinatorial polyketide biosynthesis by *de novo* design and rearrangement of modular polyketide synthase genes. Nat. Biotechnol. 23:1171-1176.

Mertz JL, Peloso JS, Barker BJ, Babbitt GE, Occolowitz JL, Simson VL, Kline RM. 1986. Isolation and structural identification of nine avilamycins. J. Antibiot. (Tokyo) 39:877-887.

Mika JT, Poolman B. 2011. Macromolecule diffusion and confinement in prokaryotic cells. Curr. Opin. Biotechnol. 22:117-126.

Minowa Y, Araki M, Kanehisa M. 2007. Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. J. Mol. Biol. 368:1500-1517.

Minton AP, Rivas G. 2011. Biochemical reactions in the crowded and confined physiological environment: Physical chemistry meets synthetic biology. In: The Minimal Cell. Dordrecht: Springer. pp. 73-89.

Mithani A, Hein J, Preston GM. 2011. Comparative analysis of metabolic networks provides insight into the evolution of plant pathogenic and nonpathogenic lifestyles in *Pseudomonas*. Mol. Biol. Evol. 28:483-499.

Mochizuki S, Hiratsu K, Suwa M, Ishii T, Sugino F, Yamada K, Kinashi H. 2003. The large linear plasmid pSLA2-L of *Streptomyces rochei* has an unusually condensed gene organization for secondary metabolism. Mol. Microbiol. 48:1501-1510.

Molbak L, Tett A, Ussery DW, Wall K, Turner S, Bailey M, Field D. 2003. The plasmid genome database. Microbiology 149:3043-3045.

Montero Llopis P, Jackson AF, Sliusarenko O, Surovtsev I, Heinritz J, Emonet T, Jacobs-Wagner C. 2010. Spatial organization of the flow of genetic information in bacteria. Nature 466:77-81.

Moon TS, Dueber JE, Shiue E, Prather KL. 2010. Use of modular, synthetic scaffolds for improved production of glucaric acid in engineered *E. coli*. Metab. Eng. 12:298-305.

Mukherji S, van Oudenaarden A. 2009. Synthetic biology: Understanding biological design from synthetic circuits. Nat. Rev. Genet. 10:859-871.

Müller U, van Assema F, Gunsior M, Orf S, Kremer S, Schipper D, Wagemans A, Townsend CA, Sonke T, Bovenberg R et al. 2006. Metabolic engineering of the *E. coli* L-phenylalanine pathway for the production of D-phenylglycine (D-phg). Metab. Eng. 8:196-208.

Mullis KB. 1990. The unusual origin of the polymerase chain reaction. Sci. Am. 262:56-61, 64-5.

Munkres J. 1957. Algorithms for the assignment and transportation problems. J. Soc. Industr. Appl. Math. 5:32-38.

Murat D, Quinlan A, Vali H, Komeili A. 2010. Comprehensive genetic dissection of the magnetosome gene island reveals the step-wise assembly of a prokaryotic organelle. Proc. Natl. Acad. Sci. U. S. A. 107:5593-5598.

Murty MN, Devi VS. 2011. Support vector machines. In: Pattern Recognition. London: Springer. pp. 147-187.

Musser JH. 2003. Carbohydrate-based therapeutics. Burger's Medicinal Chemistry, Drug Discovery and Development .

Myronovskyi M, Welle E, Fedorenko V, Luzhetskyy A. 2011. Beta-glucuronidase as a sensitive and versatile reporter in actinomycetes. Appl. Environ. Microbiol. 77:5370-5383.

Na D, Lee D. 2010. RBSDesigner: Software for designing synthetic ribosome binding sites that yields a desired level of protein expression. Bioinformatics 26:2633-2634.

Nagrath D, Avila-Elchiver M, Berthiaume F, Tilles AW, Messac A, Yarmush ML. 2010. Soft constraints-based multiobjective framework for flux balance analysis. Metab. Eng. 12:429-445.

Nakano C, Kim HK, Ohnishi Y. 2011. Identification and characterization of the linalool/nerolidol synthase from *Streptomyces clavuligerus*. Chembiochem 12:2403-2407.

Netolitzky DJ, Wu X, Jensen SE, Roy KL. 1995. Giant linear plasmids of beta-lactam antibiotic producing *Streptomyces*. FEMS Microbiol. Lett. 131:27-34.

Nett M, Ikeda H, Moore BS. 2009. Genomic basis for natural product biosynthetic diversity in the actinomycetes. Nat. Prod. Rep. 26:1362-1384.

Neumann H, Slusarczyk AL, Chin JW. 2010. *De novo* generation of mutually orthogonal aminoacyl-tRNA synthetase/tRNA pairs. J. Am. Chem. Soc. 132:2142-2144.

Neumann H, Wang K, Davis L, Garcia-Alai M, Chin JW. 2010. Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. Nature 464:441-444.

Newman DJ, Cragg GM. 2012. Natural products as sources of new drugs over the 30 years from 1981 to 2010. J. Nat. Prod. 75:311-335.

Newman DJ, Cragg GM. 2007. Natural products as sources of new drugs over the last 25 years. J. Nat. Prod. 70:461-477.

Nguyen DD, Wu CH, Moree WJ, Lamsa A, Medema MH, Zhao X, Gavilan RG, Aparicio M, Atencio L, Jackson C et al. 2013. MS/MS networking guided analysis of molecule and gene cluster families. Proc. Natl. Acad. Sci. U. S. A., in press.

Nguyen QT, Merlo ME, Medema MH, Jankevics A, Breitling R, Takano E. 2012. Metabolomics methods for the synthetic biology of secondary metabolism. FEBS Lett. 586:2177-2183.

Nguyen T, Ishida K, Jenke-Kodama H, Dittmann E, Gurgui C, Hochmuth T, Taudien S, Platzer M, Hertweck C, Piel J. 2008. Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. Nat. Biotechnol. 26:225-233.

Nierman WC, Pain A, Anderson MJ, Wortman JR, Kim HS, Arroyo J, Berriman M, Abe K, Archer DB, Bermejo C. 2005. Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. Nature 438:1151-1156.

Nieselt K, Battke F, Herbig A, Bruheim P, Wentzel A, Jakobsen OM, Sletta H, Alam MT, Merlo ME, Moore J et al. 2010. The dynamic architecture of the metabolic switch in *Streptomyces coelicolor*. BMC Genomics 11:10.

Nishida H, Ohnishi Y, Beppu T, Horinouchi S. 2007. Evolution of γ-butyrolactone synthases and receptors in *Streptomyces*. Environ. Microbiol. 9:1986.

Oberhardt MA, Palsson BØ, Papin JA. 2009. Applications of genome-scale metabolic reconstructions. Mol. Syst. Biol. 5:320.

Oberhardt MA, Puchalka J, Martins dos Santos VA, Papin JA. 2011. Reconciliation of genome-scale metabolic reconstructions for comparative systems analysis. PLoS Comput. Biol. 7:e1001116.

Oberhardt MA, Goldberg JB, Hogardt M, Papin JA. 2010. Metabolic network analysis of *Pseudomonas aeruginosa* during chronic cystic fibrosis lung infection. J. Bacteriol. 192:5534-5548.

O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. 2011. Open Babel: An open chemical toolbox. J. Cheminform 3:33-2946-3-33.

Ohnishi Y, Ishikawa J, Hara H, Suzuki H, Ikenoya M, Ikeda H, Yamashita A, Hattori M, Horinouchi S. 2008. Genome sequence of the streptomycin-producing microorganism *Streptomyces griseus* IFO 13350. J. Bacteriol. 190:4050-4060.

Olano C, Mendez C, Salas JA. 2010. Post-PKS tailoring steps in natural product-producing actinomycetes from the perspective of combinatorial biosynthesis. Nat. Prod. Rep. 27:571-616.

Oliva B, O'Neill A, Wilson JM, O'Hanlon PJ, Chopra I. 2001. Antimicrobial properties and mode of action of the pyrrothine holomycin. Antimicrob. Agents Chemother. 45:532-539.

Oliynyk M, Samborskyy M, Lester JB, Mironenko T, Scott N, Dickens S, Haydock SF, Leadlay PF. 2007. Complete genome sequence of the erythromycin-producing bacterium *Saccharopolyspora erythraea* NRRL23338. Nat. Biotechnol. 25:447-453.

Onaka H, Taniguchi S, Igarashi Y, Furumai T. 2002. Cloning of the staurosporine biosynthetic gene cluster from *Streptomyces* sp. TP-A0274 and its heterologous expression in *Streptomyces lividans*. J. Antibiot. (Tokyo) 55:1063-1071.

Orth JD, Thiele I, Palsson BØ. 2010. What is flux balance analysis? Nat. Biotechnol. 28:245-248.

Ostash B, Saghatelian A, Walker S. 2007. A streamlined metabolic pathway for the biosynthesis of moenomycin A. Chem. Biol. 14:257-267.

Ostash B, Doud EH, Lin C, Ostash I, Perlstein DL, Fuse S, Wolpert M, Kahne D, Walker S. 2009. Complete characterization of the seventeen step moenomycin biosynthetic pathway. Biochemistry 48:8830-8841.

Pagani I, Liolios K, Jansson J, Chen IA, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC. 2012. The Genomes OnLine Database (GOLD) v. 4: Status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. 40:D571-D579.

Papin JA, Price ND, Palsson BØ. 2002. Extreme pathway lengths and reaction participation in genome-scale metabolic networks. Genome Research 12:1889-1900.

Paradise EM, Kirby J, Chan R, Keasling JD. 2008. Redirection of flux through the FPP branch-point in *Saccharomyces cerevisiae* by down-regulating squalene synthase. Biotechnol. Bioeng. 100:371-378.

Paradkar AS, Aidoo KA, Jensen SE. 1998. A pathway-specific transcriptional activator regulates late steps of clavulanic acid biosynthesis in *Streptomyces clavuligerus*. Mol. Microbiol. 27:831-843.

Parsons JB, Frank S, Bhella D, Liang M, Prentice MB, Mulvihill DP, Warren MJ. 2010. Synthesis of empty bacterial microcompartments, directed organelle protein incorporation, and evidence of filament-associated organelle movement. Mol. Cell 38:305-315.

Paulsen IT, Press CM, Ravel J, Kobayashi DY, Myers GS, Mavrodi DV, DeBoy RT, Seshadri R, Ren Q, Madupu R. 2005. Complete genome sequence of the plant commensal *Pseudomonas fluorescens* pf-5. Nat. Biotechnol. 23:873-878.

Pavoine S, Baguette M, Bonsall MB. 2010. Decomposition of trait diversity among the nodes of a phylogenetic tree. Ecol. Monogr. 80:485-507.

Pelzer S, Sussmuth R, Heckmann D, Recktenwald J, Huber P, Jung G, Wohlleben W. 1999. Identification and analysis of the balhimycin biosynthetic gene cluster and its use for manipulating glycopeptide biosynthesis in *Amycolatopsis mediterranei* DSM5908. Antimicrob. Agents Chemother. 43:1565-1573.

Penn K, Jenkins C, Nett M, Udwary DW, Gontang EA, McGlinchey RP, Foster B, Lapidus A, Podell S, Allen EE et al. 2009. Genomic islands link secondary metabolism to functional adaptation in marine actinobacteria. ISME. J. 3:1193-1203.

Pfeifer BA, Khosla C. 2001. Biosynthesis of polyketides in heterologous hosts. Microbiol. Mol. Biol. Rev. 65:106-118.

Pfeifer BA, Admiraal SJ, Gramajo H, Cane DE, Khosla C. 2001. Biosynthesis of complex polyketides in a metabolically engineered strain of *E. coli*. Science 291:1790-1792.

Pharkya P, Burgard AP, Maranas CD. 2004. OptStrain: A computational framework for redesign of microbial production systems. Genome Res. 14:2367-2376.

Posfai G, Plunkett G,III, Feher T, Frisch D, Keil GM, Umenhoffer K, Kolisnychenko V, Stahl B, Sharma SS, de AM et al. 2006. Emergent properties of reduced-genome *Escherichia coli*. Science 312:1044-1046.

Prather KLJ, Martin CH. 2008. *De novo* biosynthetic pathways: Rational design of microbial chemical factories. Curr. Opin. Biotechnol. 19:468-474.

Press WH, Flannery BP, Teukolsky SA, Vetterling WT. 1992. Numerical recipes in C: The art of scientific programming. Section 10:408-412.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS ONE 5:e9490.

Price ND, Reed JL, Palsson BØ. 2004. Genome-scale models of microbial cells: Evaluating the consequences of constraints. Nat. Rev. Microbiol. 2:886-897.

Price ND, Papin JA, Schilling CH, Palsson BØ. 2003. Genome-scale microbial *in silico* models: The constraints-based approach. Trends Biotechnol. 21:162-169.

Prieto C, Garcia-Estrada C, Lorenzana D, Martin JF. 2012. NRPSsp: Non-ribosomal peptide synthase substrate predictor. Bioinformatics 28:426-427.

Puigbo P, Guzman E, Romeu A, Garcia-Vallve S. 2007. OPTIMIZER: A web server for optimizing the codon usage of DNA sequences. Nucleic Acids Res. 35:W126-31.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J et al. 2012. The Pfam protein families database. Nucleic Acids Res. 40:D290-301.

Purnick PE, Weiss R. 2009. The second wave of synthetic biology: From modules to systems. Nat. Rev. Mol. Cell Biol. 10:410-422.

Quan J, Saaem I, Tang N, Ma S, Negre N, Gong H, White KP, Tian J. 2011. Parallel on-chip gene synthesis and application to optimization of protein expression. Nat. Biotechnol. 29:449-452.

Raes J, Bork P. 2008. Molecular eco-systems biology: Towards an understanding of community function. Nat. Rev. Microbiol. 6:693-699.

Rao CV. 2012. Expanding the synthetic biology toolbox: Engineering orthogonal regulators of gene expression. Curr. Opin. Biotechnol. 23:689-694.

Rappé MS, Giovannoni SJ. 2003. The uncultured microbial majority. Annu. Rev. Microbiol. 57:369-394.

Rausch C, Hoof I, Weber T, Wohlleben W, Huson DH. 2007. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. BMC Evol. Biol. 7:78.

Rausch C, Weber T, Kohlbacher O, Wohlleben W, Huson DH. 2005. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). Nucleic Acids Res. 33:5799-5808.

Ravel J, Wellington EM, Hill RT. 2000. Interspecific transfer of *Streptomyces* giant linear plasmids in sterile amended soil microcosms. Appl. Environ. Microbiol. 66:529-534.

Ravel J, Schrempf H, Hill RT. 1998. Mercury resistance is encoded by transferable giant linear plasmids in two Chesapeake Bay *Streptomyces* strains. Appl. Environ. Microbiol. 64:3383-3388.

Revanna KV, Krishnakumar V, Dong Q. 2009. A web-based software system for dynamic gene cluster comparison across multiple genomes. Bioinformatics 25:956-957.

Rialle S, Felicori L, Dias-Lopes C, Peres S, El Atia S, Thierry AR, Amar P, Molina F. 2010. BioNetCAD: Design, simulation and experimental validation of synthetic biochemical networks. Bioinformatics 26:2298-2304.

Richardson SM, Nunley PW, Yarrington RM, Boeke JD, Bader JS. 2010. GeneDesign 3.0 is an updated synthetic biology toolkit. Nucleic Acids Res. 38:2603-2606.

Richter F, Leaver-Fay A, Khare SD, Bjelic S, Baker D. 2011. *De novo* enzyme design using Rosetta3. PLoS One 6:e19230.

Ridley DD. 2001. Introduction to structure searching with SciFinder scholar. J. Chem. Educ. 78:559.

Ro DK, Paradise EM, Ouellet M, Fisher KJ, Newman KL, Ndungu JM, Ho KA, Eachus RA, Ham TS, Kirby J et al. 2006. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature 440:940-943.

Rocha I, Maia P, Evangelista P, Vilaca P, Soares S, Pinto JP, Nielsen J, Patil KR, Ferreira EC, Rocha M. 2010. OptFlux: An open-source software platform for *in silico* metabolic engineering. BMC Syst. Biol. 4:45.

Rodrigo G, Carrera J, Jaramillo A. 2007. Asmparts: Assembly of biological model parts. Syst. Synth. Biol. 1:167-170.

Rodrigo G, Carrera J, Prather KJ, Jaramillo A. 2008. DESHARKY: Automatic design of metabolic pathways for optimal cell growth. Bioinformatics 24:2554-2556.

Rodríguez-García A, Combes P, Perez-Redondo R, Smith MC, Smith MC. 2005. Natural and synthetic tetracycline-inducible promoters for use in the antibiotic-producing bacteria *Streptomyces*. Nucleic Acids Res. 33:e87.

Rodríguez-García A, de la FA, Perez-Redondo R, Martin JF, Liras P. 2000. Characterization and expression of the arginine biosynthesis gene cluster of *Streptomyces clavuligerus*. J. Mol. Microbiol. Biotechnol. 2:543-550.

Romero J, Liras P, Martin JF. 1986. Utilization of ornithine and arginine as specific precursors of clavulanic acid. Appl. Environ. Microbiol. 52:892-897.

Roodbeen R, van Hest JC. 2009. Synthetic cells and organelles: compartmentalization strategies. Bioessays 31:1299-1308.

Röttig M, Rausch C, Kohlbacher O. 2010. Combining structure and sequence information allows automated prediction of substrate specificities within enzyme families. PLoS. Comput. Biol. 6:e1000636.

Röttig M, Medema MH, Blin K, Weber T, Rausch C, Kohlbacher O. 2011. NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity. Nucleic Acids Res. 39:W362-7.

Rückert C, Blom J, Chen X, Reva O, Borriss R. 2011. Genome sequence of *B. amyloliquefaciens* type strain DSM7(T) reveals differences to plant-associated *B. amyloliquefaciens* FZB42. J. Biotechnol. 155:78-85.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. 2000. Artemis: Sequence visualization and annotation. Bioinformatics 16:944-945.

Salanoubat M, Genin S, Artiguenave F, Gouzy J, Mangenot S, Arlat M, Billault A, Brottier P, Camus JC, Cattolico L et al. 2002. Genome sequence of the plant pathogen *Ralstonia solanacearum*. Nature 415:497-502.

Salis HM. 2011. The ribosome binding site calculator. Methods Enzymol. 498:19-42.

Salis HM, Mirsky EA, Voigt CA. 2009. Automated design of synthetic ribosome binding sites to control protein expression. Nat. Biotechnol. 27:946-950.

Salowe SP, Marsh EN, Townsend CA. 1990. Purification and characterization of clavaminate synthase from *Streptomyces clavuligerus*: An unusual oxidative enzyme in natural product biosynthesis. Biochemistry 29:6499-6508.

Santamarta I, Perez-Redondo R, Lorenzana LM, Martin JF, Liras P. 2005. Different proteins bind to the butyrolactone receptor protein ARE sequence located upstream of the regulatory *ccaR* gene of *Streptomyces clavuligerus*. Mol. Microbiol. 56:824-835.

Santoyo G, Romero D. 2005. Gene conversion and concerted evolution in bacterial genomes. FEMS Microbiol. Rev. 29:169-183.

Sattely ES, Fischbach MA, Walsh CT. 2008. Total biosynthesis: *In vitro* reconstitution of polyketide and nonribosomal peptide pathways. Nat. Prod. Rep. 25:757-793.

Saudagar PS, Survase SA, Singhal RS. 2008. Clavulanic acid: A review. Biotechnol. Adv. 26:335-351.

Scheffler RJ, Colmer S, Tynan H, Demain AL, Gullo VP. 2013. Antimicrobials, drug discovery, and genome mining. Appl. Microbiol. Biotechnol. 97:969-978.

Schellenberger J, Que R, Fleming RM, Thiele I, Orth JD, Feist AM, Zielinski DC, Bordbar A, Lewis NE, Rahmanian S et al. 2011. Quantitative prediction of cellular metabolism with constraint-based models: The COBRA toolbox v2.0. Nat. Protoc. 6:1290-1307.

Scherlach K, Hertweck C. 2009. Triggering cryptic natural product biosynthesis in microorganisms. Org. Biomol. Chem. 7:1753-1760.

Scherr N, Nguyen L. 2009. *Mycobacterium* versus *Streptomyces*--we are different, we are the same. Curr. Opin. Microbiol. 12:699-707.

Schneiker S, Perlova O, Kaiser O, Gerth K, Alici A, Altmeyer MO, Bartels D, Bekel T, Beyer S, Bode E. 2007. Complete genome sequence of the myxobacterium *Sorangium cellulosum*. Nat. Biotechnol. 25:1281-1289.

Schuetz R, Zamboni N, Zampieri M, Heinemann M, Sauer U. 2012. Multidimensional optimality of microbial metabolism. Science 336:601-604.

Schwarzer D, Finking R, Marahiel MA. 2003. Nonribosomal peptides: From genes to products. Nat. Prod. Rep. 20:275-287.

Segre D, Vitkup D, Church GM. 2002. Analysis of optimality in natural and perturbed metabolic networks. Proc. Natl. Acad. Sci. U. S. A. 99:15112-15117.

Seyed-Allaei H, Bianconi G, Marsili M. 2006. Scale-free networks with an exponent less than two. Phys. Rev. E. Stat. Nonlin. Soft Matter Phys. 73:046113.

Shen B. 2003. Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. Curr. Opin. Chem. Biol. 7:285-295.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. Nat. Biotechnol. 26:1135-1145.

Sherman DH. 2005. The lego-ization of polyketide biosynthesis. Nat. Biotechnol. 23:1083-1084.

Shoji J, Hinoo H, Katayama T, Nakagawa Y, Ikenishi Y, Iwatani K, Yoshida T. 1992. Structures of new peptide antibiotics, plusbacins A1-A4 and B1-B4. J. Antibiot. (Tokyo) 45:824-831.

Shou W, Ram S, Vilar JM. 2007. Synthetic cooperation in engineered yeast populations. Proc. Natl. Acad. Sci. U. S. A. 104:1877-1882.

Silakowski B, Kunze B, Müller R. 2001. Multiple hybrid polyketide synthase/non-ribosomal peptide synthetase gene clusters in the myxobacterium *Stigmatella aurantiaca*. Gene 275:233-240.

Sirikantaramas S, Yamazaki M, Saito K. 2007. Mechanisms of resistance to self-produced toxic secondary metabolites in plants. Phytochem. Rev. 7:467-477.

Smanski MJ, Peterson RM, Rajski SR, Shen B. 2009. Engineered *Streptomyces platensis* strains that overproduce antibiotics platensimycin and platencin. Antimicrob. Agents Chemother. 53:1299-1304.

Smith D. 2003. Culture collections over the world. Int. Microbiol. 6:95-100.

Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. 2011. Cytoscape 2.8: New features for data integration and network visualization. Bioinformatics 27:431-432.

Soh KC, Hatzimanikatis V. 2010. DREAMS of metabolism. Trends Biotechnol. 28:501-508.

Starcevic A, Zucko J, Simunkovic J, Long PF, Cullum J, Hranueli D. 2008. ClustScan: An integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and *in silico* prediction of novel chemical structures. Nucleic Acids Res. 36:6882-6892.

Steen EJ, Kang Y, Bokinsky G, Hu Z, Schirmer A, McClure A, Del Cardayre SB, Keasling JD. 2010. Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. Nature 463:559-562.

Stevens DC, Henry MR, Murphy KA, Boddy CN. 2010. Heterologous expression of the oxytetracycline biosynthetic pathway in *Myxococcus xanthus*. Appl. Environ. Microbiol. 76:2681-2683.

Stover BC, Müller KF. 2010. TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics 11:7.

Straight PD, Fischbach MA, Walsh CT, Rudner DZ, Kolter R. 2007. A singular enzymatic megacomplex from *Bacillus subtilis*. Proc. Natl. Acad. Sci. U. S. A. 104:305-310.

Streit WR, Schmitz RA. 2004. Metagenomics--the key to the uncultured microbes. Curr. Opin. Microbiol. 7:492-498.

Strobel T, Al-Dilaimi A, Blom J, Gessner A, Kalinowski J, Luzhetska M, Puhler A, Szczepanowski R, Bechthold A, Rückert C. 2012. Complete genome sequence of *Saccharothrix espanaensis* DSM 44229(T) and comparison to the other completely sequenced *Pseudonocardiaceae*. BMC Genomics 13:465-2164-13-465.

Stubblefield BA, Howery KE, Islam BN, Santiago AJ, Cardenas WE, Gilbert ES. 2010. Constructing multispecies biofilms with defined compositions by sequential deposition of bacteria. Appl. Microbiol. Biotechnol. 86:1941-1946.

Tae H, Sohng JK, Park K. 2009. MapsiDB: An integrated web database for type I polyketide synthases. Bioprocess. Biosyst. Eng. 32:723-727.

Tahlan K, Anders C, Jensen SE. 2004. The paralogous pairs of genes involved in clavulanic acid and clavam metabolite biosynthesis are differently regulated in *Streptomyces clavuligerus*. J. Bacteriol. 186:6286-6297.

Tahlan K, Park HU, Jensen SE. 2004. Three unlinked gene clusters are involved in clavam metabolite biosynthesis in *Streptomyces clavuligerus*. Can. J. Microbiol. 50:803-810.

Tahlan K, Park HU, Wong A, Beatty PH, Jensen SE. 2004. Two sets of paralogous genes encode the enzymes involved in the early stages of clavulanic acid and clavam metabolite biosynthesis in *Streptomyces clavuligerus*. Antimicrob. Agents Chemother. 48:930-939.

Tahlan K, Anders C, Wong A, Mosher RH, Beatty PH, Brumlik MJ, Griffin A, Hughes C, Griffin J, Barton B. 2007. 5S clavam biosynthetic genes are located in both the clavam and paralog gene clusters in *Streptomyces clavuligerus*. Chem. Biol. 14:131-142.

Takahashi H, Kumagai T, Kitani K, Mori M, Matoba Y, Sugiyama M. 2007. Cloning and characterization of a *Streptomyces* single module type non-ribosomal peptide synthetase catalyzing a blue pigment synthesis. J. Biol. Chem. 282:9073-9081.

Takano E. 2006. γ-Butyrolactones: *Streptomyces* signalling molecules regulating antibiotic production and differentiation. Curr. Opin. Microbiol. 9:287-294.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56:564-577.

Tamsir A, Tabor JJ, Voigt CA. 2011. Robust multicellular computing using genetically encoded NOR gates and chemical 'wires'. Nature 469:212-215.

Tang SY, Fazelinia H, Cirino PC. 2008. AraC regulatory protein mutants with altered effector specificity. J. Am. Chem. Soc. 130:5267-5271.

Temme K, Zhao D, Voigt CA. 2012. Refactoring the nitrogen fixation gene cluster from *Klebsiella oxytoca*. Proc. Natl. Acad. Sci. U. S. A. 109:7085-7090.

Temme K, Hill R, Segall-Shapiro TH, Moser F, Voigt CA. 2012. Modular control of multiple pathways using engineered orthogonal T7 polymerases. Nucleic Acids Res. 40:8773-8781.

Thiele I, Palsson BØ. 2010. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat. Protoc. 5:93-121.

Thomson AW, Bonham CA, Zeevi A. 1995. Mode of action of tacrolimus (FK506): Molecular and cellular mechanisms. Ther. Drug Monit. 17:584-591.

Thornton JW. 2004. Resurrecting ancient genes: Experimental analysis of extinct molecules. Nat. Rev. Genet. 5:366-375.

Tobias NJ, Doig KD, Medema MH, Chen H, Haring V, Moore R, Seemann T, Stinear TP. 2013. Complete genome sequence of the frog pathogen *Mycobacterium ulcerans* ecovar Liflandii. J. Bacteriol. 195:556-564.

Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, van Hijum SA. 2012. Data mining in the life sciences with random forest: A walk in the park or lost in the jungle? Brief. Bioinform. 14:315-326.

Trefzer A, Pelzer S, Schimana J, Stockert S, Bihlmaier C, Fiedler HP, Welzel K, Vente A, Bechthold A. 2002. Biosynthetic gene cluster of simocyclinone, a natural multihybrid antibiotic. Antimicrob. Agents Chemother. 46:1174-1182.

Tseng HC, Prather KL. 2012. Controlled biosynthesis of odd-chain fuels and chemicals via engineered modular metabolic pathways. Proc. Natl. Acad. Sci. U. S. A. 109:17925-17930.

Tyo KE, Kocharin K, Nielsen J. 2010. Toward design-based engineering of industrial microbes. Curr. Opin. Microbiol. 13:255-262.

Udwary DW, Zeigler L, Asolkar RN, Singan V, Lapidus A, Fenical W, Jensen PR, Moore BS. 2007. Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. Proc. Natl. Acad. Sci. U. S. A. 104:10376-10381.

Umezawa H, Ueda M, Maeda K, Yagishita K, Kondo S, Okami Y, Utahara R, Osato Y, Nitta K, Takeuchi T. 1957. Production and isolation of a new antibiotic: kanamycin. J. Antibiot. (Tokyo) 10:181-188.

Ussery DW, Wassenaar TM, Borini S. 2008. Microbial communities: Core and pan-genomics. In: Anonymous Computing for Comparative Microbial Genomics. London: Springer. pp. 213-228.

van den Berg MA, Albang R, Albermann K, Badger JH, Daran JM, Driessen AJ, Garcia-Estrada C, Fedorova ND, Harris DM, Heijne WH et al. 2008. Genome sequencing and analysis of the filamentous fungus *Penicillium chrysogenum*. Nat. Biotechnol. 26:1161-1168.

van Dongen S, Abreu-Goodger C. 2012. Using MCL to extract clusters from networks. Methods Mol. Biol. 804:281-295.

van Hijum SA, Medema MH, Kuipers OP. 2009. Mechanisms and evolution of control logic in prokaryotic transcriptional regulation. Microbiol. Mol. Biol. Rev. 73:481-509.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al. 2001. The sequence of the human genome. Science 291:1304-1351.

Villalobos A, Ness JE, Gustafsson C, Minshull J, Govindarajan S. 2006. Gene designer: A synthetic biology tool for constructing artificial DNA segments. BMC Bioinformatics 7:285.

Viollier PH, Minas W, Dale GE, Folcher M, Thompson CJ. 2001a. Role of acid metabolism in *Streptomyces coelicolor* morphological differentiation and antibiotic biosynthesis. J. Bacteriol. 183:3184-3192.

Viollier PH, Nguyen KT, Minas W, Folcher M, Dale GE, Thompson CJ. 2001b. Roles of aconitase in growth, metabolism, and morphological differentiation of *Streptomyces coelicolor*. J. Bacteriol. 183:3193-3203.

Vo TD, Greenberg HJ, Palsson BØ. 2004. Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. J. Biol. Chem. 279:39532-39540.

Volff JN, Altenbuchner J. 1998. Genetic instability of the *Streptomyces* chromosome. Mol. Microbiol. 27:239-246.

Wagner A. 2007. Energy costs constrain the evolution of gene expression. J Exp. Zool. B Mol. Dev. Evol. 308:322-324.

Waisvisz J, Van Der Hoeven M, Van Peppen J, Zwennis W. 1957. Bottromycin. I. A new sulfur-containing antibiotic. J. Am. Chem. Soc. 79:4520-4521.

Walsh CT, Fischbach MA. 2010. Natural products version 2.0: Connecting genes to molecules. J. Am. Chem. Soc. 132:2469-2493.

Wang K, Neumann H, Peak-Chew SY, Chin JW. 2007. Evolved orthogonal ribosomes enhance the efficiency of synthetic genetic code expansion. Nat. Biotechnol. 25:770-777.

Ward JM, Janssen GR, Kieser T, Bibb MJ, Buttner MJ, Bibb MJ. 1986. Construction and characterisation of a series of multi-copy promoter-probe plasmid vectors for *Streptomyces* using the aminoglycoside phosphotransferase gene from Tn5 as indicator. Mol. Gen. Genet. 203:468-478.

Watanabe K, Hotta K, Praseuth AP, Koketsu K, Migita A, Boddy CN, Wang CC, Oguri H, Oikawa H. 2006. Total biosynthesis of antitumor nonribosomal peptides in *Escherichia coli*. Nat. Chem. Biol. 2:423-428.

Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, van der Voort M, Pogliano K, Gross H, Raaijmakers JM et al. 2012. Mass spectral molecular networking of living microbial colonies. Proc. Natl. Acad. Sci. U. S. A. 109:E1743-52.

Watrous JD, Dorrestein PC. 2011. Imaging mass spectrometry in microbiology. Nat. Rev. Microbiol. 9:683-694.

Watve MG, Tickoo R, Jog MM, Bhole BD. 2001. How many antibiotics are produced by the genus *Streptomyces*? Arch. Microbiol. 176:386-390.

Weber T. 2013. *In silico* tools for the analysis of antibiotic biosynthetic pathways. Int. J. Med. Microbiol., in press.

Weber T, Rausch C, Lopez P, Hoof I, Gaykova V, Huson DH, Wohlleben W. 2009. CLUSEAN: A computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. J. Biotechnol. 140:13-17.

Weber W, Daoud-El Baba M, Fussenegger M. 2007. Synthetic ecosystems based on airborne inter- and intrakingdom communication. Proc. Natl. Acad. Sci. U. S. A. 104:10435-10440.

Weeding E, Houle J, Kaznessis YN. 2010. SynBioSS designer: A web-based tool for the automated generation of kinetic models for synthetic biological constructs. Brief. Bioinform. 11:394-402.

Weeks AM, Chang MC. 2011. Constructing *de novo* biosynthetic pathways for chemical synthesis inside living cells. Biochemistry 50:5404-5418.

Wehmeier UF, Piepersberg W. 2004. Biotechnology and molecular biology of the alpha-glucosidase inhibitor acarbose. Appl. Microbiol. Biotechnol. 63:613-625.

Weininger D. 1988. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. 28:31-36.

Weissman KJ, Müller R. 2008. Protein-protein interactions in multienzyme megasynthetases. Chembiochem. 9:826-848.

Whitaker WR, Davis SA, Arkin AP, Dueber JE. 2012. Engineering robust control of two-component system phosphotransfer using modular scaffolds. Proc. Natl. Acad. Sci. U. S. A. 109:18090-18095.

Widenbrant EM, Tsai HH, Chen CW, Kao CM. 2007. *Streptomyces coelicolor* undergoes spontaneous chromosomal end replacement. J. Bacteriol. 189:9117-9121.

Wilkinson B, Micklefield J. 2009. Biosynthesis of nonribosomal peptide precursors. Methods Enzymol. 458:353-378.

Willett P, Barnard JM, Downs GM. 1998. Chemical similarity searching. J. Chem. Inf. Comput. Sci. 38:983-996.

Williams DE, Craig M, Holmes CF, Andersen RJ. 1996. Ferintoic acids A and B, new cyclic hexapeptides from the freshwater cyanobacterium *Microcystis aeruginosa*. J. Nat. Prod. 59:570-575.

Wingler LM, Cornish VW. 2011. Reiterative recombination for the *in vivo* assembly of libraries of multigene pathways. Proc. Natl. Acad. Sci. U. S. A. 108:15135-15140.

Winter JM, Behnken S, Hertweck C. 2011. Genomics-inspired discovery of natural products. Curr. Opin. Chem. Biol. 15:22-31.

Wishart DS. 2009. Computational strategies for metabolite identification in metabolomics. Bioanalysis 1:1579-1596.

Withers ST, Keasling JD. 2007. Biosynthesis and engineering of isoprenoid small molecules. Appl. Microbiol. Biotechnol. 73:980-990.

Wohlleben W, Mast Y, Muth G, Rottgen M, Stegmann E, Weber T. 2012. Synthetic biology of secondary metabolite biosynthesis in actinomycetes: Engineering precursor supply as a way to optimize antibiotic production. FEBS Lett. 586:2171-2176.

Wong FT, Khosla C. 2012. Combinatorial biosynthesis of polyketides--a perspective. Curr. Opin. Chem. Biol. 16:117-123.

Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ et al. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 462:1056-1060.

Wu G, Culley DE, Zhang W. 2005. Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* and the implications for their metabolism. Microbiology 151:2175-2187.

Wu M, Eisen JA. 2008. A simple, fast, and accurate method of phylogenomic inference. Genome Biol. 9:R151.

Wu W, Leblanc SK, Piktel J, Jensen SE, Roy KL. 2006. Prediction and functional analysis of the replication origin of the linear plasmid pSCL2 in *Streptomyces clavuligerus*. Can. J. Microbiol. 52:293-300.

Wu X, Roy KL. 1993. Complete nucleotide sequence of a linear plasmid from *Streptomyces clavuligerus* and characterization of its RNA transcripts. J. Bacteriol. 175:37-52.

Xu D. 2009. Computational methods for protein sequence comparison and search. Curr. Protoc. Protein Sci. 56:2.1.1-2.1.27.

Yadav G, Gokhale RS, Mohanty D. 2009. Towards prediction of metabolic products of polyketide synthases: An *in silico* analysis. PLoS. Comput. Biol. 5:e1000351.

Yadav G, Gokhale RS, Mohanty D. 2003. Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. J. Mol. Biol. 328:335-363.

Yamada T, Letunic I, Okuda S, Kanehisa M, Bork P. 2011. iPath2.0: Interactive pathway explorer. Nucleic Acids Res. 39:W412-W415.

Yamasaki M, Kinashi H. 2004. Two chimeric chromosomes of *Streptomyces coelicolor* A3 (2) generated by single crossover of the wild-type chromosome and linear plasmid SCP1. J. Bacteriol. 186:6553-6559.

Yanai K, Murakami T, Bibb M. 2006. Amplification of the entire kanamycin biosynthetic gene cluster during empirical strain improvement of *Streptomyces kanamyceticus*. Proc. Natl. Acad. Sci. U. S. A. 103:9661-9666.

Yepes A, Rico S, Rodríguez-García A, Santamaria RI, Diaz M. 2011. Novel two component systems implied in antibiotic production in *Streptomyces coelicolor*. PLoS ONE 5: e19980.

Yilmaz P, Gilbert JA, Knight R, Amaral-Zettler L, Karsch-Mizrachi I, Cochrane G, Nakamura Y, Sansone SA, Glockner FO, Field D. 2011. The Genomic Standards Consortium: Bringing standards to life for microbial ecology. ISME J. 5:1565-1567.

You L, Cox RS,III, Weiss R, Arnold FH. 2004. Programmed population control by cell-cell communication and regulated killing. Nature 428:868-871.

Zakrzewski P, Medema MH, Gevorgyan A, Kierzek AM, Breitling R, Takano E. 2012. MultiMetEval: Comparative and multi-objective analysis of genome-scale metabolic models. PLoS One 7:e51511.

Zaslaver A, Mayo AE, Rosenberg R, Bashkin P, Sberro H, Tsalyuk M, Surette MG, Alon U. 2004. Just-in-time transcription program in metabolic pathways. Nat. Genet. 36:486-491.

Zazopoulos E, Huang K, Staffa A, Liu W, Bachmann BO, Nonaka K, Ahlert J, Thorson JS, Shen B, Farnet CM. 2003. A genomics-guided approach for discovering and expressing cryptic metabolic pathways. Nat. Biotechnol. 21:187-190.

Zeigler DR. 2011. The genome sequence of *Bacillus subtilis* subsp. *spizizenii* W23: Insights into speciation within the *B. subtilis* complex and into the history of *B. subtilis* genetics. Microbiology 157:2033-2041.

Zelyas NJ, Cai H, Kwong T, Jensen SE. 2008. Alanylclavam biosynthetic genes are clustered together with one group of clavulanic acid biosynthetic genes in *Streptomyces clavuligerus*. J. Bacteriol. 190:7957.

Zerikly M, Challis GL. 2009. Strategies for the discovery of new natural products by genome mining. Chembiochem. 10:625-633.

Zha W, Rubin-Pitel SB, Shao Z, Zhao H. 2009. Improving cellular malonyl-CoA level in *Escherichia coli* via metabolic engineering. Metab. Eng. 11:192-198.

Zhang H, Boghigian BA, Armando J, Pfeifer BA. 2011. Methods and options for the heterologous production of complex natural products. Nat. Prod. Rep. 28:125-151.

Zhang W, Li Y, Tang Y. 2008. Engineered biosynthesis of bacterial aromatic polyketides in *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. 105:20683-20688.

Zhao XQ, Gust B, Heide L. 2010. S-adenosylmethionine (SAM) and antibiotic biosynthesis: Effect of external addition of SAM and of overexpression of SAM biosynthesis genes on novobiocin production in *Streptomyces*. Arch. Microbiol. 192:289-297.

Ziemert N, Jensen PR. 2012. Phylogenetic approaches to natural product structure prediction. Methods Enzymol. 517:161.

Ziemert N, Podell S, Penn K, Badger JH, Allen E, Jensen PR. 2012. The natural product domain seeker NaPDoS: A phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. PLoS ONE 7:e34064.

Zotchev SB, Sekurova ON, Katz L. 2012. Genome-based bioprospecting of microbes for new therapeutics. Curr. Opin. Biotechnol. 23:941-947.

Zucko J, Long PF, Hranueli D, Cullum J. 2012. Horizontal gene transfer and gene conversion drive evolution of modular polyketide synthases. J. Ind. Microbiol. Biotechnol. 39:1541-1547.

English summary for the non-specialist

Bacteria, fungi and plants produce large numbers of small molecules, a range of which is applied medicinally as antibiotics, chemotherapeutics, cholesterol-lowering agents and immunosuppressants. We developed a new strategy for the discovery of novel drugs from such molecules, which exploits the enormous amount of hereditary information from these life forms that has recently become available through new technologies. This allowed the exploration of the systems that produce molecules with medicinal potential throughout thousands of life forms. Moreover, the thesis outlines a novel approach to characterize the molecules produced by these systems in large numbers, by redesigning the systems from scratch on a computer and inserting them into engineered bacteria.

Many medicines originate from small but complex chemical molecules made by bacteria, fungi and plants, which are called *secondary metabolites* (introduced in Chapter 1). They are called 'secondary', because unlike 'primary' metabolites, the molecules are not absolutely necessary for survival: instead, in their native environment, they perform important accessory functions such as communication, warfare, and nutrient uptake. Secondary metabolites are constructed by the action of small 'machines', called *enzymes*, which are proteins that help direct chemical reactions to transform one molecule into another.

There is still a great need to discover more of these molecules, specifically antibiotics. At increasing rates, bacteria are becoming resistant to the antibiotics that are currently used. There are even forms of tuberculosis that cannot be treated anymore, because the TB bacterium has become resistant to all antibiotics on the market.

Naturally, one way to find new drugs is to look at more bacteria and fungi and to see whether they produce any drug-like molecules that have not been identified before. After all, different bacteria and fungi produce different molecules: many carry unique enzymes to fabricate molecules not found in any other *organisms* (life forms).

Hereditary information

Why do different life forms produce different molecules in the first place? This has to do with the hereditary information that they carry on their DNA. DNA is a molecule that is in some way like a large string of letters, each represented by a chemical group called a 'base.' All the letters of this code together form things that are in a way like 'sentences': *genes*. A gene contains information that can be read to construct, for example, an enzyme. In this way, DNA functions as the 'programming language' of life. The entire collection of genes and other DNA elements of an organism is called a *genome*, and it comprises all the instructions necessary to build the organism and to make the molecules that the organism needs to survive.

In the last decades, technologies have been developed to read this DNA code. In 1995, the first complete genome of a bacterium was read, and in 2001, two competing scientific teams both presented an initial complete 'text' of the human genome. Since that time, the technology has

rapidly improved. By now, the entire genomes of thousands of bacteria, fungi, plants and animals have been read. Their DNA information is stored on computers and is freely available for access via a web browser.

Gene clusters

The total DNA information of just one of these life forms usually contains thousands of genes. Recent research showed that the DNA of many bacteria contains multiple 'biosynthetic' gene clusters: groups of genes working together that are physically clustered together on the DNA, which encode machinery to make secondary metabolites—the broad class of molecules mentioned earlier that could potentially be used as drugs. In Chapter 8 of this thesis, we even showed that the genome of one bacterium contains around 50 of these gene clusters. Interestingly, many more gene clusters were found than secondary metabolites had been discovered for this bacterium, and the same had been found for other bacteria previously. This indicates more and more strongly that under standard laboratory conditions, only a fraction of these machineries is actually active, and leads to the production of metabolites (Chapter 9). Hence, there is clear potential to use this genetic information to find novel molecules that can be developed into drugs.

However, it is still very challenging to identify those genes and molecules that truly have the potential to deliver novel drugs, because the amount of information to search through is enormous. If one would put all the available DNA letters together in books, they would comprise an entire library. Therefore, the only way to make sense of all this information is through the power of computers. This allows one to decode the programming language of life and search through it, just as if it were a Google search.

New software tools

The main program that we wrote (Chapters 2 and 3), antiSMASH, allows users to simply open a web browser and upload the entire genome of a certain bacterium or fungus. Then, the program will find all the genes present in this genome and search through them to find those gene clusters that encode enzymes, the 'machines' of life, that are likely to be able to construct secondary metabolites. Thus, the program can quickly present the user with a (nearly) complete overview of all the secondary metabolites that the bacterium or fungus can potentially produce, given what is written in its DNA. For some metabolites, it can even already give a prediction of what the chemical structure of the molecule will look like, based on predictions of the functionality of the enzymes that build it.

Besides antiSMASH, we also wrote additional computer programs that compare sets of genes with all the genes found in other genomes (Chapter 4), that match the genes with experimentally detected molecules (Chapter 5) and that simulate how efficient the enzymes can produce metabolites given their context in a certain life form (Chapter 6).

Mapping the haystack

With these tools, we could then automatically read the 'programs' of life to make chemical compounds. Given that DNA data from thousands of organisms have recently become available, this allowed us to develop a radically new strategy to find novel drugs from secondary metabolites. Traditionally, researchers would often look for new molecules in quite randomly chosen bacteria, which is a bit like trying to find a few needles in a haystack. Instead, we decided to first map the whole haystack and look globally for things that look like needles, before zooming in further: we performed a global analysis of all the genes that encode the production of interesting molecules in thousands of organisms, and analyzed these data with computer programs (Chapter 10). This made it possible to see which organisms are likely to produce the most unique or most unexplored molecules, which gives laboratory researchers a much better chance of truly finding something novel.

Engineering life

It is a great asset to be able to read life's programming language. Logically, the question arises: can we also write it? This very question recently inspired a group of biological engineers to launch a new field: synthetic biology. Synthetic biology represents the idea that by writing DNA codes one can engineer and build life. For example, one can construct a new bacterium that attacks cancer cells, or engineer bacteria to monitor when food goes off. But of course one can also engineer bacteria to produce secondary metabolites. That is why we outlined a strategy to exploit this: if bacteria are first modified to become very proficient in making these molecules, one could then simply take gene clusters encoding the production of a range of interesting molecules and one by one insert these into their DNA (Chapters 11 and 12). This would allow the characterization of lots of compounds, even if the organism they come from cannot be grown in the laboratory or normally does not produce the molecule at high enough levels.

At first, this 'plug-and-play' strategy would allow for the writing and testing of existing systems throughout the biosphere. Later on, when the action and interplay of enzymes can be modeled more precisely with computers, it could even become possible to build novel designer molecules entirely from scratch, by bringing together the right combination of enzymes and encoding this in a working stretch of DNA (Chapter 7). Thus, I predict that the combination of computational techniques and synthetic biology will radically transform the way we find new drugs from microbes (Chapter 13). Hopefully, this will empower the identification of many novel antibiotics as well as novel drugs to fight cancer.

Nederlandse samenvatting voor de leek

Bacteriën en schimmels produceren allerlei kleine moleculen, waarvan een groot aantal medisch toegepast wordt als antibiotica, chemotherapeutica, cholesterolverlagende middelen of immunosuppressiva. Dit proefschrift beschrijft een nieuwe strategie om medicijnen te ontwikkelen uit deze moleculen, door gebruik te maken van de enorme hoeveelheid erfelijke informatie van deze levensvormen die door nieuwe technologieën beschikbaar is geworden. Dit maakte het mogelijk om in duizenden levensvormen tegelijk te zoeken naar systemen die moleculen met medische potentie produceren. Om deze systemen en de moleculen die erdoor gemaakt worden vervolgens ook daadwerkelijk uit te testen stellen we een innovatieve benadering voor: de systemen kunnen op de computer herschreven worden en vervolgens kunstmatig ingebracht worden in speciaal daarvoor ontworpen bacteriën.

Veel medicijnen zijn afkomstig van kleine maar complexe moleculen die gemaakt worden door bacteriën, schimmels en planten. Deze moleculen noemen we ook wel *secundaire metabolieten* (zie Hoofdstuk 1). Ze worden 'secundair' genoemd, omdat de levensvorm die ze produceert ze in tegenstelling tot 'primaire' metabolieten niet absoluut nodig heeft om te overleven. In plaats daarvan vervullen ze in hun natuurlijke omgeving allerlei secundaire functies, zoals communicatie, oorlogsvoering en de opname van voedingsstoffen. Secundaire metabolieten worden gebouwd door kleine 'machientjes' die *enzymen* genoemd worden. Dit zijn eiwitten die ervoor zorgen dat een chemische reactie het ene molecuul in het andere omzet.

Het is hard nodig dat er meer van dit soort moleculen ontdekt worden, en dan met name antibiotica: bacteriën zijn in toenemende mate resistent tegen de op dit moment gebruikte middelen. Er zijn zelfs vormen van tuberculose die niet meer behandeld kunnen worden, omdat de tuberculosebacterie resistent is geworden tegen alle verkrijgbare antibiotica.

Eén manier om nieuwe medicijnen te vinden is natuurlijk om naar meer verschillende bacteriën en schimmels te kijken om te zien of zij misschien moleculen maken die nog niet eerder geïdentificeerd zijn. Want verschillende levensvormen maken verschillende moleculen: veel bacteriën hebben zelfs de beschikking over unieke enzymen die moleculen maken die in geen andere levensvorm gevonden worden.

Erfelijke informatie

Waarom is het eigenlijk zo dat verschillende levensvormen verschillende moleculen maken? Dit heeft te maken met de erfelijke informatie die opgeslagen ligt in hun DNA. DNA is een molecuul dat in zekere zin lijkt op een lange ketting van letters, die elk vertegenwoordigd worden door een chemische groep die een 'base' heet. Alle letters uit deze code vormen samen een soort 'zinnen', die *genen* genoemd worden. Een gen bevat de informatie om bijvoorbeeld een enzym te produceren. Het DNA fungeert in feite als de 'programmeertaal' van het leven. De gehele verzameling aan genen en andere DNA-elementen van een levensvorm wordt een *genoom* genoemd. Zo'n genoom bevat in principe alle instructies om de levensvorm te bouwen en de moleculen te maken die het nodig heeft

om te overleven.

In de laatste decennia zijn er technologieën ontwikkeld die het mogelijk maken om de DNA-code te lezen. In 1995 werd het eerste complete genoom van een bacterie gelezen, en in 2001 presenteerden twee afzonderlijke wetenschappelijke teams een eerste vrijwel volledige 'tekst' van het menselijk genoom. Sinds die tijd heeft de techniek zich rap verder ontwikkeld. Op dit moment zijn de genomen van duizenden bacteriën, schimmels, planten en dieren gelezen. De DNA-informatie hieruit wordt bewaard op computers en is vrij toegankelijk via internet.

Genclusters

Alle DNA-informatie van slechts één levensvorm bevat meestal duizenden verschillende genen. Recent onderzoek heeft aangetoond dat het DNA van veel bacteriën verscheidene 'biosynthetische' genclusters bevat: groepen van bij elkaar in de buurt liggende genen die samenwerken om de machinerie te produceren om secundaire metabolieten te maken — de eerdergenoemde klasse moleculen die potentie heeft om als medicijnen gebruikt te worden. In Hoofdstuk 8 van dit proefschrift laten we zelfs zien dat het genoom van één specifieke bacterie wel vijftig van deze genclusters bevat. Interessant genoeg vonden we veel meer genclusters dan dat er secundaire metabolieten ontdekt zijn voor deze bacterie. Hetzelfde was eerder ook gevonden voor andere bacteriën. Dit suggereert dat slechts een fractie van deze machinerieën actief metabolieten produceert onder standaardlaboratoriumomstandigheden (Hoofdstuk 9). Daarom biedt de genetische informatie duidelijk nieuwe mogelijkheden om moleculen te ontdekken die kunnen leiden tot de ontwikkeling van medicijnen.

Desalniettemin is het nog steeds een grote uitdaging om genen en moleculen te vinden die ook daadwerkelijk kans maken om het tot een medicijn te schoppen, mede omdat de hoeveelheid informatie die men moet doorzoeken enorm is. Als je alle beschikbare DNA-letters zou samenvoegen in boeken, zou je een hele bibliotheek kunnen vullen. Daarom kan deze informatie alleen effectief aangewend worden door gebruik te maken van computers. Die maken het mogelijk om de programmeertaal van het leven te ontcijferen en er als het ware 'doorheen te googelen'.

Nieuwe computerprogramma's

Het belangrijkste computerprogramma dat we hebben geschreven (Hoofdstuk 2 en 3), antiSMASH, maakt het voor gebruikers mogelijk om simpelweg via een webbrowser het gehele genoom van een bacterie of schimmel te uploaden. Het programma zoekt vervolgens alle genen waaruit het genoom is opgebouwd en doorzoekt deze om genclusters te vinden die de code bevatten om enzymen te produceren die secundaire metabolieten maken. Op deze manier kan het programma de gebruiker razendsnel een (vrijwel) compleet beeld geven van alle secundaire metabolieten die een bacterie of schimmel lijkt te kunnen produceren op grond van zijn DNA. Voor sommige metabolieten kan antiSMASH zelfs al een voorspelling doen van hoe de chemische structuur van het molecuul eruit ziet, gebaseerd op voorspellingen van de functies van de enzymen die het lijken te bouwen.

Naast antiSMASH hebben we nog een aantal computerprogramma's geschreven. Eén vergelijkt de

bovengenoemde ‘genclusters’ met genen die in andere genomen gevonden worden (Hoofdstuk 4), een ander probeert voor experimenteel gedetecteerde moleculen het bijpassende gencluster te vinden (Hoofdstuk 5), en een laatste simuleert hoe efficiënt enzymen zijn in het produceren van bepaalde metabolieten op grond van hun context binnen de levensvorm waarin ze werkzaam zijn (Hoofdstuk 6).

Hooiberg

Met deze computertechnieken konden wij vervolgens automatisch de ‘programma’s’ lezen die levensvormen gebruiken om chemische stofjes in elkaar te zetten. Omdat er recentelijk DNA-informatie van duizenden levensvormen beschikbaar is gekomen, konden we dit vervolgens aanwenden om een vernieuwende strategie ontwikkelen om nieuwe secundaire metabolieten te ontdekken die potentie hebben om als medicijn te dienen. Traditioneel zochten onderzoekers vaak naar nieuwe moleculen in relatief willekeurig gekozen bacteriën, wat een beetje is alsof je op zoek gaat naar een handjevol spelden in een grote hooiberg. Om het zoekproces efficiënter te maken, besloten we om als het ware eerst de hele hooiberg in kaart te brengen en te kijken waar in die hooiberg zich de meeste ‘speld-achtige’ structuren bevinden: we deden een globale analyse van alle genen die coderen voor de productie van interessante moleculen in duizenden levensvormen tegelijk, en analyseerden die data met de computer (Hoofdstuk 10). Dit maakte het mogelijk om te zien welke organismen waarschijnlijk de meest unieke en onontdekte moleculen maken. Met behulp van deze informatie hebben onderzoekers in het laboratorium veel meer kans om echt nieuwe stofjes te vinden die potentie te hebben om doorontwikkeld te worden tot medicijn.

Het leven bouwen

Het is van groot nut om de programmeertaal van het leven te kunnen lezen. Logischerwijs komt dan de vraag op: kunnen we deze programmeertaal ook zelf schrijven? Deze vraag heeft recent een groep biologisch technologen ertoe gedreven om een nieuw onderzoeksveld op te starten: synthetische biologie. Synthetische biologie gaat uit van het idee dat we als mensen zelf het leven kunnen ontwerpen en bouwen door DNA-code te schrijven. Zo zou je bijvoorbeeld een bacterie kunnen bouwen die kankercellen aanvalt, of een bacterie die het opmerkt als voedsel niet meer houdbaar is. Maar natuurlijk kan je ook een bacterie bouwen die secundaire metabolieten produceert. Daarom hebben we een strategie ontwikkeld om daar gebruik van te maken: als bacteriën eerst aangepast worden om heel goed te worden in het maken van zulke moleculen, kan je vervolgens een aantal met de computer gevonden genclusters nemen die voor de productie van interessante moleculen lijken te coderen en deze inbouwen in het DNA van de aangepaste bacterie (Hoofdstuk 11 en 12). Dit maakt het mogelijk om heel veel verschillende stofjes uit allerlei verschillende levensvormen te bestuderen, zelfs al weten we niet hoe we de levensvorm in het laboratorium kunnen laten groeien of zelfs al produceert het normaal gesproken maar een minuscule hoeveelheid van het molecuul.

Allereerst zou deze ‘plug-en-play’-strategie het mogelijk maken om bestaande productiesystemen voor moleculen uit de biosfeer te herschrijven en uit te testen, en te zien wat voor functie de

moleculen hebben. En zodra de werking en interactie van enzymen preciezer met computers gemodelleerd kan worden, wordt het misschien zelfs mogelijk om nieuwe moleculen van de grond af aan te ontwerpen, door de juiste combinatie van enzymen bij elkaar te brengen en dit alles te coderen in een functioneel stuk DNA (Hoofdstuk 7). Mijn voorspelling is dat aldus de combinatie van computertechnieken en synthetische biologie de manier waarop medicijnen ontdekt worden uit microben drastisch zal veranderen (Hoofdstuk 13). Hopelijk maakt dit de weg vrij voor de ontdekking van vele nieuwe antibiotica en medicijnen tegen kanker.

Acknowledgements

I am very glad that doing a PhD does not mean that one locks oneself up in a vacuum for four years only to come out again at the end with a thesis. Even though in theory a PhD thesis is 'mainly' the product of one person, it's incredible how many people are actually involved. I have a lot of reasons to be thankful, and a lot of people to be thankful towards.

Eriko and Rainer, of course none of this would have been possible without you. You were always there to provide input and ideas, to answer e-mails even while traveling, and to critically review manuscripts.

Eriko, thank you for introducing me into the world of *Streptomyces* biology. Thank you for your guidance and your critical thinking. And thank you for allowing me the freedom to follow my own ideas and interests within the framework of the project.

Rainer, thank you for helping me keep focus on the things that matter. Thanks for your original methodological ideas that greatly strengthened this thesis. And thanks for helping me become a better writer through your creative feedback.

Ritsert and Lubbert, I would also like to specifically thank you for hosting me in your departments during the work. Thanks for your positive encouragement, and for helping to create an environment that stimulates creativity.

Much of what makes me excited about the thesis comes from collaborations. I have learned a lot from many collaborators, and they also brought in expertise that I would never have had the time or ability for to acquire myself.

Kai and Tilmann, thanks for the great collaboration on antiSMASH. Kai, thanks for sharing the rough month of December 2010 with me, when we worked day and night to get our software finished. Doing it together made it fun. Thanks also for teaching me to think not just as a scientist, but also as a software engineer. Marc and Christian, thanks for the nice collaboration on NRPSPredictor2.

Michael, thanks for hosting me in San Francisco and for giving me the chance not only to get to know this exciting city but also letting me part of the exciting science in your group. Thanks for your confidence and for the fun we had with analyzing data and writing it up into an exciting story. Peter, the same counts for you of course: it was great fun to work together with you, and this thesis benefited greatly from your insights, your methodological rigor and your creative ideas. Mohamed, Laurens, Jan, Brianna, Mao, thanks for making the time at UCSF a great experience.

Roel, Ulrike, Wilbert, Axel, Marco and Liang, thanks for sharing the exciting data on *Streptomyces clavuligerus*, and for sharing your insights on industrial practice as it happens at DSM, which made the analyses and discussions on these data much more relevant and realistic.

Pieter and Don, thanks for allowing me to contribute to your exciting development of peptidogenomic approaches. I really enjoy working with you.

And I should not forget Tim, Nick, Natalie, Victor, Albert and Andrzej. Thanks for the work we could do together.

During my project I had the pleasure of working with a number of excellent and inspiring students. Thank you very much for sharing in the excitement of this wonderful topic, Konrad, Renske, Piotr, Ronald and Yared!

Elena, we shared in the joy of being part of two research groups. When I sometimes did not know where I really belonged, you were always there. Thanks for all the great conversations that we had, with or without coffee, on science or on life in general.

Lara, Ana, HaJö, and Veronica: thanks for always making the lunch breaks at Zernike a time to look forward to. And thanks for being great 'strepto' colleagues. The same of course counts for Marco, Davide, Wouter, Dennis, Ilse en Nai-Hua in the 'early years' and also for Mirjan, Xianfeng, Andrij, Thai, Santiago and Scott. Many thanks also go out to all the other MicPhys members for providing a great work environment: Mark, Marta, Joana, Alicia, Vincent, Evelien, Hans, Sander, Maarten, Jelle, Geralt, Pieter, Alle, Robert, Theo, Jolanda, Justyna, Gerda, and anybody that I might have forgotten: thanks!

Bea, Manon and Klazien, thanks for the great work that you did in the secretariats; you were my navigation system through the maze of bureaucracy that the university can sometimes be.

Tauqeer, we worked together a lot in the first year. Thanks for helping me get into the project and helping me to start enjoying it. Andris and Richard, thanks for being great office mates in the first year. I always got new energy just from listening to your jokes or your lovely quasi-cynicism.

Yang, Danny, Bruno, Frank, Maria, René, Minh Anh, Pariya, Lionel, Joeri, Morris and Peter: thanks for the inspiring environment that the GBiC offered. From the 'tax office' in Haren to the deep dungeons of the Linnaeusborg, it was always a great place to work.

I would also like to thank my reading committee for taking the time and effort to read through my thesis. Jörn Piel, Peter Leadlay and Marcel Reinders: many thanks!

Ik wil mijn ouders bedanken dat zij mij altijd gesteund en gestimuleerd hebben, wat er absoluut toe heeft bij gedragen dat ik nu sta waar ik sta, met een proefschrift dat af is. Bedankt voor wie jullie zijn! Ook de rest van de familie en vrienden ontzettend bedankt voor jullie interesse en steun. Daarbij wil ik in het bijzonder Wart noemen: bedankt voor je prachtige design van het proefschrift!

José, jou ben ik het meeste dankbaar van alle mensen die ik noem. Door het feit dat ik elke dag thuis kon komen en me door jou onvoorwaardelijk geliefd mocht weten, kon ik vol vertrouwen en met

veel plezier mijn onderzoek aanpakken en tegelijkertijd ook mijn werk in het juiste perspectief plaatsen. Dankjewel voor wie je voor me bent! En ik bedank ook het wondertje in je buik dat licht en vreugde heeft gebracht tijdens het laatste jaar…

Tenslotte wil ik God bedanken, die de Schepper is van al het moois dat ik heb mogen bestuderen. Wat een ongelooflijk adembenemend complexe en prachtige natuur heeft hij gemaakt! Ik voel me bevoorrecht dat ik mij er elke dag opnieuw over mag verwonderen…

Curriculum vitae

Marinus Hendrik (Marnix) Medema was born on January 24th, 1986, in Vaassen, the Netherlands. After obtaining his B.Sc. in Biology *cum laude* at the Radboud University Nijmegen in 2006, he moved to Groningen to participate in the Top Master Programme in Biomolecular Sciences at the University of Groningen. After research projects with Prof. Dr. Jan Kok in Groningen and Dr. Marc Strous in Nijmegen, he obtained his M.Sc. degree *cum laude* in 2008. From 2008 to 2009, he did a study year in philosophy and theology at the University of Tilburg, aided by a scholarship from the Radboudstichting. From August 2009 to July 2013, he performed his PhD research at the University of Groningen under the supervision of Prof. Dr. Eriko Takano and Prof. Dr. Rainer Breitling, with a project on the genome mining and synthetic biology implementation of secondary metabolite biosynthetic gene clusters from microbes. During this period, he spent five months at the group of Dr. Michael Fischbach at the University of California, San Francisco, to jointly develop new methods and initiate a global analysis of secondary metabolite biosynthesis throughout all sequenced bacterial genomes. Since August 2013, Marnix is working as a postdoctoral researcher at the group of Frank Oliver Glöckner at the Max Planck Institute for Marine Microbiology in Bremen, supported by a Rubicon grant from the Dutch Science Foundation NWO.